Development of a prediction model for pipeline failure probability based on learning from past incidents and pipeline specific data using artificial neural network (ANN)

March 2022

CAAP Final Report

Date of Report: February 23, 2022

Prepared for: *U.S. DOT Pipeline and Hazardous Materials Safety Administration*

Contract Number: Cooperative Agreement #693JK31850011CAAP

Project Title: Development of a prediction model for pipeline failure probability based on learning from past incidents and pipeline specific data using artificial neural network (ANN)

Prepared by: Noor Quddus (Principal Investigator)
Mary Kay O'Connor Process Safety Center (MKOPSC)
Texas A&M Engineering Experiment Station (TEES)

Contact Information:

Table of Contents

List of Figures

List of Tables

Executive Summary

The United States Pipeline and Hazardous Materials Safety Administration (PHMSA) has been gathering incident data for last thirty-five years. The rules and criteria have been changed a few times with the development of new findings and necessity. At present, PHMSA is collecting over 600 data fields for every reported incident. Some of these incidents are required to go through full investigations based on specific criteria. They provide more detail insight about the incidents especially the detail cause(s) behind the incident. They also offer a way forward to improve and prevent any future incidents of the same or similar kind. However, none of the systems or process of gathering incident data or incident investigation is flawless. They must go through a continuous improvement process. New pipeline technologies, increase in commodity transported and hence the number of incidents, and novel data analytic techniques especially the machine learning and artificial intelligence tools necessitate the continuous improvement process to keep going. In the current study, natural language processing (NLP) and artificial neural network (ANN) have been employed to analyze past pipeline incident data and develop models to predict future pipeline incidents.

PHMSA collects incidents data particularly in three areas based on the commodity transported: hazardous liquid (HL), gas transmission and gathering (GTG), and gas distribution (GD). The data files are available for three different timelines (1986-2001/2004), (2002/2004-2009), and (2010-present) because of rules changes of reporting criteria. Current study considered HL database for year (2010-present) for appropriate data size and data quality. Total number of incidents in this dataset is 3,755. Because of changes in the reporting system, data from different time periods are not consistent, and hence, only the newest set of data was selected. On the same note, incident investigation reports prepared after 2010 and related to HL were considered for further analysis (total number of reports considered is 44).

A comprehensive literature review related to causal analysis of pipeline failures was conducted. Primary focus of the published articles includes identifying the major causes of failure, failure trends, and their relationship with other parameters. According to the literature, researchers used data from various sources including PHMSA datasets. As it necessitates, causal structures as reported in other incident data sources, such as National Energy Board, Canada (NEB) and European Gas Pipeline Incident Data Group (EGIG), were also studied. Analysis of the obtained

incident data and causal-consequence relationship showed that factors that influenced the pipeline incidents can be classified into three groups:

- Causal factors
    - direct causes – mapped causes and sub-causes as coined by PHMSA
    - Seven standard causes: corrosion, material/ weld/ equipment failure, excavation damage, incorrect operation, natural force damage, other outside force damage, all other causes
    - Reporting multiple causes is not allowed by PHMSA
- Background factors
    - All relevant information
    - Such as pipe diameter, wall thickness, material of construction, year of installation, commodity transported, operating pressure, coating material
    - PHMSA collection of background information is comprehensive
- Underlying factors
    - They are the root causes
    - Such as poor maintenance protocol, inadequate supervision, faulty design
    - PHMSA does not gather such information in a structured manner

There is no recommended restructure found for incident investigation reports. They often more focused on only technical issues and failed to identify the underlying causes especially management or organizational issues. Since there is not structured guideline to do it, it is often difficult to summarize them. Another limitation is the small number of publicly available incident investigation reports (less than hundred for all categories).

To gather more information than that are available in incident records obtained from the databases, two sources have been used: first, the incident narratives from the incident database, and second, the available incident investigation reports. Three NLP and text mining techniques have been explored to extract useful information from the two datasets: K-means clustering, topic modeling, and co-occurrence network. K-means clustering and co-occurrence network have been applied to incident narratives, and topic modeling and co-occurrence to investigation reports. Since data was structured and organized against direct causes, K-means clustering did not produce very good results. It also indicates the secondary causes or underlying causes not consistent. In other words, any significant secondary pattern was absent. On the contrary, both co-occurrence network

and topic modeling identified words that appeared again and again in close proximity. For instance, they were able to identify the words that appeared near to a certain word say corrosion. Both unsupervised (the word corrosion not given) and supervised (the word corrosion was fed as seed word) options have been explored. Both methods found very promising in extracting background and underlying factor. A validation exercise has been conducted by manually checking the narratives where the techniques were suggestive. For further study, supervised technique around a specific direct cause or development of a multi-step NLP techniques would be beneficial.

An integrated framework for risk prediction using ANN and Bayesian model has been used to predict a corrosion-induced incident in the onshore HL pipeline. The causal and consequence estimation models have been developed utilizing the ANN technique, and the probability estimation model utilizes the Bayesian analysis. The ANN model utilized 70 data fields from the incident record and compress them into 26 using process knowledge resulting in higher information density. ANN model was validated with reasonable accuracy for several consequence categories. The Bayesian analysis model performance is also tested for the prediction of the probability of the incident. Utilizing the proposed framework including ANN models for cause and consequence prediction and Bayesian analysis for probability prediction, the risk of a corrosion-induced pipeline incident was predicted given the current condition of pipelines. This shows the strength of the proposed framework to predict the risk of corrosion-induced pipeline incidents and can further be extended to pipeline incidents caused by other causes such as excavation, natural forces, etc.

Summary of Accomplishments

Technical Publications and Presentations

Peer-reviewed journal publications:

- Kumari, P., Halim, S.Z., Kwon, J.S., and Quddus, N., An integrated risk prediction model for corrosion-induced pipeline incidents using artificial neural network and Bayesian analysis, submitted to *Process Safety and Environmental Protection* (under review).
- Liu, G., Boyd, M., Yu, M., Halim, S. Z., & Quddus, N. (2021). Identifying causality and contributory factors of pipeline incidents by employing natural language processing and text mining techniques. *Process Safety and Environmental Protection, 152*, 37-46.
- Halim, S. Z., Yu, M., Escobar, H., & Quddus, N. (2020). Towards a causal model from pipeline incident data analysis. *Process Safety and Environmental Protection*, *143*, 348-360.

Peer-reviewed conference presentations:

- Noor Quddus, Guanyang Liu, Mason Boyd, Pallavi Kumari, and Syeda Zohra Halim, "Automated Workflow for Causation Analysis in Pipeline Incidents: An AI-based Approach" presented virtually in PRCI Virtual Research Exchange 2021 conference (Virtual).
- Noor Quddus, Guanyang Liu, Mason Boyd, Pallavi Kumari, and Syeda Zohra Halim, "How Well Can We Predict Causes behind the Pipeline Incidents?" presented virtually in 2021 AIChE Spring Conference and 17th Global Congress on Process Safety (Virtual).
- Pallavi Kumari and Noor Quddus, Causation Analysis of Pipeline Incidents Using Artificial Neural Network, Presented virtually at MKOPSC Process Safety International Symposium, October 2020 (Virtual).
- Guanyang Liu, Mason Boyd, Pallavi Kumari, Syeda Zohra Halim, and Noor Quddus, An Intelligent Learning Framework for Analysis of Pipeline Incident Investigation Reports, presented virtually in MKOPSC Process Safety International Symposium 2020, College Station, October 2020.
- Pallavi Kumari and Noor Quddus, Causation Analysis of Pipeline Incidents Using Artificial Neural Network, presented virtually in 2020 AIChE Spring Meeting & 16th Global Congress on Process Safety (virtual)
- Guanyang Liu, Mason Boyd, and Noor Quddus, Extracting Causal Relations from Incident Reports: A Natural Language Processing and Topic Modeling, presented virtually in 2020 AIChE Spring Meeting & 16th Global Congress on Process Safety
- Guanyang Liu, Mason Boyd, and Noor Quddus, Analysis of Pipeline Incident Data and Investigation Reports Using Natural Language Processing (NLP), presented virtually in Hazards30,

18-20 May, Manchester, UK

Non-peer-reviewed Presentations:

- Liu, G., Kumari, P., Boyd, M., Quddus, N., and Holste, J. Development of a Prediction Model for Pipeline Failure Probability based on Learnings from Past Incidents and Pipeline Specific Data using Artificial Neural Network (ANN), Poster presented at the PHMSA R&D Forum, February 2020
- Regular updates have been provided to the member at the MKOPSC Steering Committee meetings which held six times a year.

Student Engagement

- Guanyang Liu, (Ph. D. student, Chemical Engineering) – 2nd and 3rd year of the project
- Pallavi Kumari, (Ph. D. student, Chemical Engineering) – 2nd and 3rd year of the project
- Harold Escober, (Ph. D. student, Chemical Engineering) – 1st year of the project
- Mengxi Yu, (Ph. D. student, Chemical Engineering) (completed) – 1st and 3rd year of the project
- Mason Boyd, Undergraduate student researcher, Chemical Engineering (completed) – Most part of the project

1.0 INTRODUCTION

On-site inspection, laboratory analysis such as chemical analysis, metallography, mechanical testing such as failure tests, fatigue test, and other data analysis are conducted to determine when a pipeline will likely fail if exposed to a specific condition. Diverse factors including operational and environmental conditions, natural calamities, manufacturing defects, and even deficiencies in management's attempt to maintain the integrity can simultaneously influence pipeline integrity and cause early failure. It is difficult, if not impossible, to deterministically predict failures arising from such factors due to the variety and number of such factors. For better assessment and prediction of pipeline failures, these factors which can contribute to pipeline failure and their corresponding contributions needs to be learnt and implemented from past incidents.

Once an incident occurs due to pipeline failure, they are reported as per regulatory requirement and incident investigation are conducted for a few to determine what caused the incident and recommendations are provided so that proper measure can be taken to prevent future recurrences. Over the last 20 years, more than 11 thousand pipeline incidents have been reported to PHMS (Pipeline and Hazardous Materials Safety Administration 2019). If the causes behind these incidents can be identified appropriately, it can act as a vast source of knowledge. If the learnings of what went wrong in these incidents are used to understand what can go wrong in the future, then it is possible to obtain a more compete hazard analysis and better failure prediction of an installed pipeline. Unfortunately, neither the learnings/ recommendations made after an incident are always applied in practice, nor are such learnings remembered for use in future hazard analysis or risk/failure assessment. Usually, the task of hazard identification in the industry is left to a team of experts who use their experience to predict what can go wrong and how. Anything beyond their experience is not captured in any hazard identification or incident investigation process. Any factor that the experts are unaware of or have forgotten or not considered that it had caused incidents in the past will not be considered in a failure analysis. This can lead to a repetition of a past incident. With the reported 11,000+ incidents that resulted in more than 300 deaths and more than 7 billion dollars' worth damage in the past 20 years, it is apparent that incidents keep happening (Pipeline and Hazardous Materials Safety Administration 2019) and not enough learning and understanding exists to stop them. If these incidents are studied in-depth, it is likely that one will identify similarities in terms of existing root causes or contributing factors behind several of them.

Indeed, the fallacy of a complete hazard identification can be seen in several studies comparing the causes of incident that occurred with scenarios that had previously been identified in accident analysis studies. At the latest European Loss Prevention symposium, Taylor (2016) , a life-long risk analysis specialist, looked back at his own risk assessment results of 92 quantitative risk analysis (Q(Esmaeili and Hallowell 2012)RAs) of 429 plant units in 36 years. So far, 26 major hazard accidents occurred. He concluded that 20% was due to missing scenarios that had not been predicted. In the US, based on analysis of incident investigations of US Chemical Safety Board, Kaszniak (2010), followed by Baybutt (2016) found that in quite a few cases, none of the lessons learnt from previous incidents had been applied to the PHAs and often, the recommendations, such as application of correct safeguards, had not been placed into practice.

As Hollnagel (2017) mentions, accident investigation and risk assessments are two sides of the same coin in that they consider the same set of events or phenomena either retrospectively (after they have happened) or prospectively (before they have happened). It is rather easier to determine in retrospect what causes led to an incident rather than predict what can go wrong that will lead to an incident in the future. It is also easier to focus on technical issues that can lead to a failure (and these are easier to determine from laboratory experiments) rather than determining the human and organizational contributions to the causes. To achieve completeness of hazard analysis and predict failure of pipelines, it is essential to consider all learnings obtained from root cause analysis of past incidents and take proper mitigative measures based on the learnings to eliminate the causes.

Once root causes behind failures become available, it should be utilized to prevent similar incidents from occurring. Pipelines can run for thousands of miles and failure incident data are usually available in large numbers. When such large number of incidents is investigated, many root causes can also suffice. Attempting to utilize these root causes to predict future failures can be a daunting task since upfront, it may not be possible to identify the relation between the failure and the cause and the relation between the causes themselves.

A complete root cause analysis leads to identification of the deficiencies that contributed to the incident. As such, most of these deficiencies point to organizational limitations that an investigation team attempts to uncover so as to provide better recommendations. A team unravels several root causes that interacted together to cause the incident. Pipelines constructed and maintained by any organization have numerous technical, operations, human and organizational

factors interacting with each other. Factors such as training of workforce, budget allocation for maintenance, availability of resources are factors that can appear as root causes behind an incident. Such causes are non-technical in nature, and it is very difficult to determine how they will contribute to failure. As Perrow (2011) mentions, in a complex system, their interactions give rise to numerous non-linear, tightly coupled cause-effect relationships that are difficult to predict. Non-linear relationships arise when the effect of interaction of several direct and indirect causes may be something different from their simple summation. For a pipeline too, operation may be influenced by the various constituents of a socio-technical system, such as human factors (such as how an operator conducts maintenance operations) or organizational factors (such as training provided on time to maintenance personnel) and non-linear relationships among the constituent exist naturally.

In the past, attempts have been made to learn from previous incidents, some of which try to cover all the different constituent or aspects of a system. From incident investigation practice, long checklists of cue words in a taxonomy structure, preferably computerized, have been developed so that none of the causes of failure are overlooked. Taylor (2017) summarized how despite checks, reviews, and HAZOPs, errors still creep into designs. If these lists are not comprehensive it will lead to incomplete assessment. It will probably never be sufficient to cover the near-endless variety of possible causes either. A new method helping to learn from past incidents is DyPASI (Dynamic Procedure for Atypical Scenarios Identification) developed by Paltrinieri et al. (2013). The method uses data mining similarity algorithm applied to incident data bases (such as French ARIA, European eMARS, and near-misses reports and risk studies) to extract cases that have resemblance to the plant at hand. This is followed by a prioritization of results based on relevance through a similarity score to help with hazard identification. This was further modified (Paltrinieri et al. 2014) to estimate updated probability of incident scenarios by means of Bayesian inference. Rathnayaka et al. (2011) developed a predictive accident propagation model based on fault and event trees as part of an approach called the System Hazard Identification, Prediction and Prevention (SHIPP) methodology. Here also, recorded incidents are used to predict future ones. A generic incident process sequence is defined from initiation to final consequence via event tree, taking into account the various types of barriers, including prevention, dispersion, ignition, escalation, human factor, and management and organizational barrier. Fault trees are drawn for each barrier and use generic data at the basic events. With likelihood data of

major incidents being rate, this method makes use of data related to observed plant disturbances, such as unusual behavior, near misses, mishaps, to update older data in the event tree using Bayesian theorem to predict probabilities of future minor and major incidents. Quite a few other similar works that have their basis on Bayesian network have been developed (Kalantarnia et al. 2009, Xin et al. 2017, Meng et al. 2019). However, Bayesian Networks are acyclic graphs, hence feedback from an effect node back to an earlier cause level is not possible: in BN an effect cannot influence its cause. For this reason, relying only on using incident precursor data, as done in SHIPP methodology, does not allow identification of factors that have the largest potential to cause an incident. In addition, once a Bayesian network is set up, incorporation of a new cause (node) or changes in the way a system interacts cannot be modeled. In that sense, dynamic modeling, to capture the effects of changes in a system, is not obtained. Bayesian network, being dependent on cause-effect relationships, also has limitations in capturing non-linear interactions that arise when multiple factors contribute to failure. Ferjencik (Ferjencik 2011, Ferjencik 2014) attempted to overcome the problems of linear cause-effect models by combining the advantages of a predefined tree Root Cause Analysis (RCA), with those of Systems-Theoretic Accident Model and Processes (STAMP). STAMP identifies direct and indirect causality using a systems approach (Leveson 2016). In that, it can capture non-linear interaction of various factors which are otherwise not seen in the traditional cause-effect linear relationships shown by predefined trees. Ferjencik's (Ferjencik 2011) work resulted in an integrated method called IPICA (Integrated Procedure for Incident Cause Analysis). Ferjencik (Ferjencik 2011) states the limitations of the predefined root cause tree, though, by quoting the adage:" What You Look For Is What You Find". This bias leads to incompleteness and limited depth. It can also be argued what exactly is a root cause, and the RCA technique is not suitable in case of non-linear and dynamic system behavior. Therefore, a deeper investigation using STAMP is called for. The method appeared to be rather tiresome in use, mainly because for STAMP, much information must be collected and conducting STAMP is laborious. Ferjencik (Ferjencik 2014) tried to simplify the method as IPICA_Lite, while retaining as much as possible of the original IPICA. In IPICA_Lite (Ferjencik 2014) STAMP is replaced by suggestions for improvements made to root cause analysis which include identification of causes as deficiencies in the system and grouping processes connected to causes into a hierarchy. Thus, through proper root causes analysis, the prediction attained from predefined trees is improved.

The methods for implementing information regarding past failures have a common in their inability to capture the nonlinear interaction of causes that contribute to the failure of pipelines. Perhaps IPICA, using STAMP attempts to capture the managerial and the societal aspects but as shown, the method itself is tedious and its later version IPICA Lite reverts back to using the simple linear relationship among causes. Yet, for developing a model that considers such non-linear relationships, one must examine the total enterprise/organization and its culture and not only the circumstances and causation of the accident.

It is thus of great importance to understand how different factors can come together and influence the pipeline integrity. Such information can be combined with laboratory analysis and findings from inspection to better predict when and how a pipeline will fail.

1.1 Research Objectives

To develop a knowledge-based predictive model to assess pipeline failure it would require

a. Learning about causes behind pipeline failure: Conducting root cause analysis of past incidents to identify those factors that have to potential to contribute to failure. The findings are to be specific to the extent that they can be applied into a predictive model.
b. Implementation of learning to predict failure: Utilizing the learnings about contributing factors behind pipeline failure to develop a predictive model that monitors current existing conditions to determine dynamic failure probability of pipeline so that the factors can be tackled before they lead to failure.

1.2 Proposed Framework

A framework for developing a model from root cause failure analysis of past pipeline incidents has been proposed based on the artificial neural network (ANN). ANN offers great potential for the development of a monitoring system based on past records while overcoming the limitations of the past attempts. The proposal presented development of an ANN model for the prediction of failure of pipelines based on findings from investigation of past pipeline failure, both in real life operation as well as in laboratory experiments. The suitability of ANN for this purpose lies in its ability to do the following (Sidarta et al. 2017):

- learn from past records to produce a predictive model
- model complex non-linear behavior that may exist in any socio-technical system,
- recognize or classify patterns in behavior and interaction of various contributing factors,
- tolerate noises and deal with large data.

They are particularly useful when there is no prior knowledge about how the variables interact since ANN models develop an understanding of the relations based on information provided during training. Thus, past findings can be utilized to train the ANN to recognize the relations between variables (Figure 1.1).

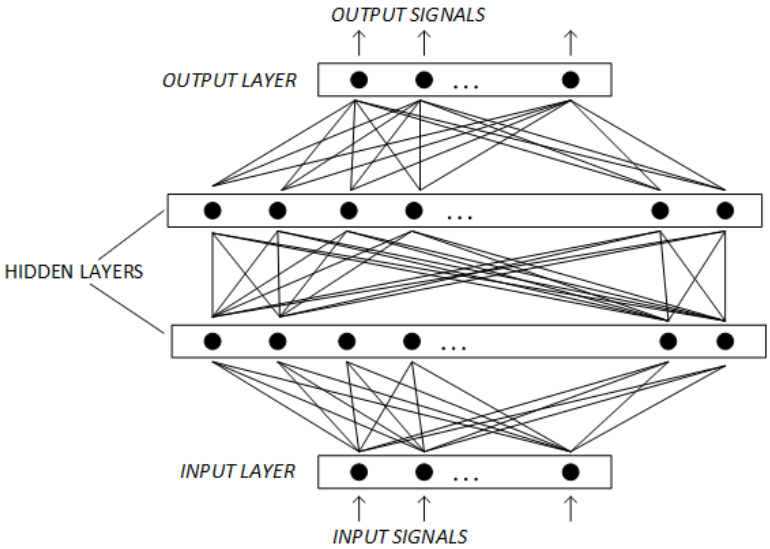

Figure 1.1 A generalized neural network, adopted from (Sidarta et al. 2017)

Artificial Neural Networks are based on the mimicry of biological neural networks that have neurons as unit processors, arranged in a multilayered structure that processes input data through hidden layers that contain stored information to help in determining the proper output. Fig 1 shows a generalized neural network model. Input signals received by neurons at the input layer propagate through the hidden layers to reach the output layers. Neurons at each layer are connected to all neurons at the next layer with a connection weight or value that is determined from past information (during training). The input signal to a neuron i (x_i) is converted to an output signal (Zhang et al.) based on a sigmoidal activation function. Between layers, the connection of neuron i to a neuron j in the next layer carries a certain weight (w_{ij}) that depends on the summation of the signals coming from all the neurons in the previous layer (Yegnanarayana 2009, Sidarta et al. 2017).

For a given input to the neurons at the input level, an output is generated as signals pass through the network. This output will be compared with the actual output for the given input and an error between the actual output and the generated output will be determined. This error is back propagated through the network to modify the connection weights. There are many developed

methods by which this modification can be obtained (*e.g.* Delta-Bar-Delta-Bar method (Ochiai and Usui 1993)).

In recent years, several attempts had been made to use ANN models to predict pipeline conditions. They are as follows:

- A model developed for sewer pipeline that utilizes information related to age, material of construction, length, diameter, depth and slope of pipes and their measured conditions to predict what condition may exist for a given pipe section (Najafi and Kulandaivel 2005).
- A model to predict burst pressure was developed using ANN after assessing the failure behavior of pipeline due to interacting defect depth (Xu et al. 2017)
- A model was developed to predict the remaining useful life of pipeline using ANN which is developed by Levenberg-Marquardt backpropagation methodology (Zangenehmadar and Moselhi 2016).
- A BP neural network is developed to identify the crack in pipeline after quantification of crack geometry (Liu et al. 2017)
- An ANN model was developed using inspection data which can predict the future condition of pipeline

These models were developed based on data encompassing inspection, crack measurements, defect depths, pipeline properties, failure behaviors *etc*. given as input to ANN. However, none of the methods had used any non-technical (human or organizational) factors such as the effect of improper maintenance or maintenance backlog and thus, does not allow adoption for implementation of findings from root cause failure analysis.

The study looks into overcoming the limitations of the past attempts by suggesting a framework for implementation of findings from past incidents along with other data sources such as those from inspection records and laboratory tests (Figure 1.2) to develop a predictive model that accounts for contribution to failure by both technical and non-technical factors.

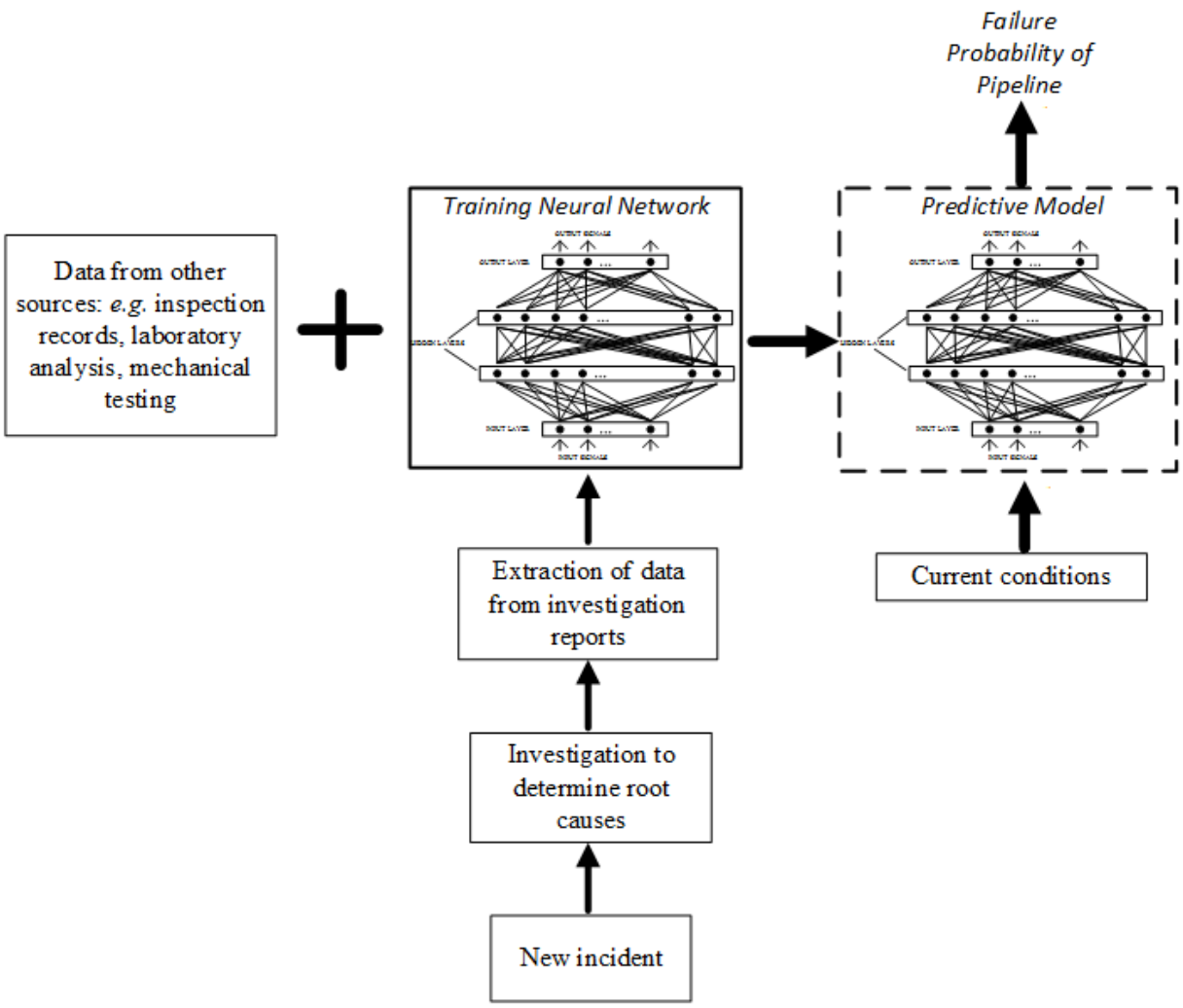

Figure 1.2 Framework for learning from past incidents to predict pipeline failure probability

1.3 Research Tasks

The framework brings forth the tasks required to meet the challenges towards application of artificial neural network (ANN) to ensure pipeline integrity. These are outlined and explained next.

1. Development of methodology for creating root cause analysis reports
2. Selection of training samples and development of the learning algorithm and validation of resulting model and utilization for prediction

This report describes the status of the pipeline incident data and incident investigation report in Chapter 2. The outline and results of the two research tasks are presented in the Chapter 3 and Chapter 4. Chapter 5 concludes with a few recommendations and future direction of research.

2.0 CURRENT STATUS OF PIPELINE INCIDENT DATA

2.1 Background

Worldwide energy demand grew by 2.3% in 2018 with natural gas emerging as the fuel of choice, showing the biggest gains and accounting for 45% of the rise in energy consumption (International Energy Agency 2019). A worldwide wave of pipeline construction activity has been driven by this continuing global shift towards natural gas. In particular, the United States led the global increase for the first time in 20 years; the rapid growth of shale production, the lifting of an oil export ban, and the predicted growth in global LNG demand led to a massive infrastructure development, including oil and LNG export terminals, and the pipeline capacity to supply them (Awalt 2019). The U.S. pipeline network which consists of two-third of world's pipeline mileage, transports almost all the natural gas produced and used in the country, as well as over 90% of crude oil and refined petroleum products. There has been a 44% increase in transportation through pipelines in the last five years (Allison and Mandler 2018, American Petroleum Institute and Associattion of Oil Pipelines 2019, Central Intelligence Agency 2019, Wikipedia 2019). The US pipeline network however is no different from the pipelines of the rest the world when it comes to pipeline incidents. Figure 2.1 shows a few well-known pipeline incidents from across the world since 2010 (Abraham 2019, Wikipedia 2019), including data of lives the claimed and the spillage they caused. It includes some major US pipeline incidents such as Kalamazoo River pipeline leak spilling 20,000 barrels of oil (National Transportation Safety Board 2012), San Bruno pipeline explosion killing 8 people (National Transportation Safety Board 2011), and Keystone pipeline spillage (Ramírez-Camacho et al.) releasing 5,000 barrels of crude oil (National Transportation Safety Board 2018). The occurrence of such major incidents at frequent intervals worldwide indicates that the challenges involving safe oil and gas transportation via pipelines are still magnanimous.

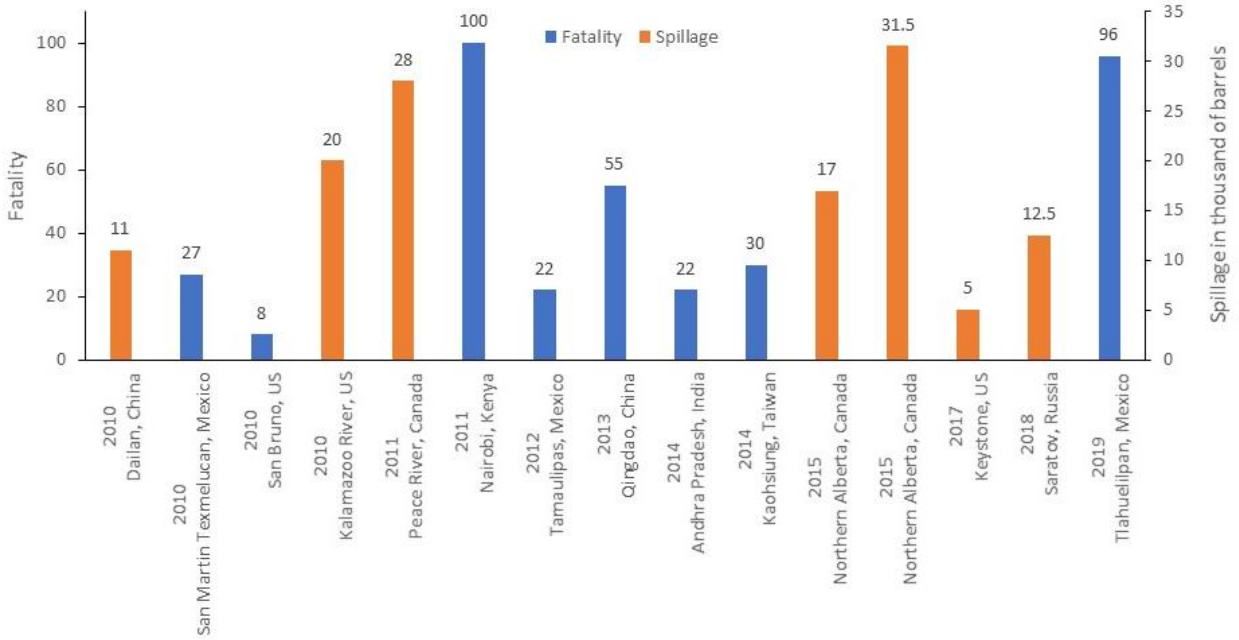

Figure 2.1 A few major pipeline incidents across the world with fatality or spillage

The Pipeline and Hazardous Materials Safety Administration (Pipeline and Hazardous Materials Safety Administration), which is responsible for safe operation of the pipeline network in the United States, publishes pipeline incident data regularly on their website (Pipeline and Hazardous Materials Safety Administration 2019). A set of safety performance parameters used by PHMSA is summarized in Table 2.1 and the corresponding trend lines are shown in Figure 2.2. These parameters (number of incidents, significant incident, fatality, injury, asset damage, and spillage) are considered *lagging indicators*, which is "a retrospective set of metrics that are based on incidents that meet the threshold of severity that should be reported as part of the industry-wide process safety metric" (Center for Chemical Process Safety 2011). The variation of parameter values especially number of injuries and spillage amount appears random as there can be multiple reasons behind it. They are used in calculating failure rates and might be useful to predict a future incident; but they do not provide much insight on how to improve the current performance or reduce the chances of future incidents. However, the sheer number of incidents, their trends and causes identified indicate that there exist ample opportunities for improvement. In the past, this has drawn a lot of researchers to this domain of pipeline research.

Table 2.1 Pipeline incident summary of hazardous liquid system in the US for the last 9 years

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
All incidents	586	588	571	618	706	712	632	647	636	5696
Significant incidents	264	285	255	303	302	329	309	302	291	2640
Fatality	22	13	12	9	19	11	16	20	8	130
Injury	108	55	57	44	95	48	87	38	90	622
Damage ($ in million)	1693	426	230	369	321	350	377	321	1050	5137
Spillage (‘000 barrels)	41.5	36.7	60.4	14.8	25.6	47.5	16.9	37.4	44.5	325

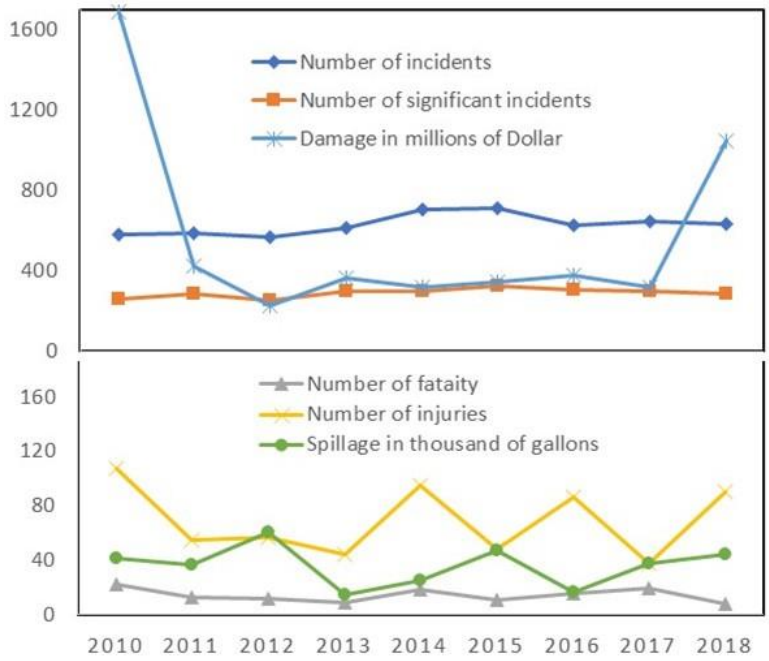

Figure 2.2 Trendlines for pipeline incident parameters observed by PHMSA

A summary of literature on cause analysis of pipeline incidents is shown in Table 2.2. The table contains the causal factors that the studies considered, background factors that were identified as having some association with the causal factors, and the source of data where the causes were reported (such as US DOT, CONCAWE, EGIG, and UKOPA). Here, the ***causal factor*** is defined as “a major unplanned, unintended contributor to an incident (a negative event or undesirable condition), that if eliminated would have either prevented the occurrence of the incident, or reduced its severity or frequency” (Center for Chemical Process Safety 2019). Thus, causal factors are those that are directly responsible for causing an incident and these factors may act alone or

with other factors to cause the incident or worsen the impact. The studies reported corrosion, mechanical failure, third-party damage, outside force etc. as causal factors. A causal factor like corrosion may contribute to an incident along, for example by causing pipe damage leading to a leakage, or it may weaken the pipeline and with slight third-party damage, may together cause a pipeline to break. A few studies have also linked other parameters such as commodity transported, pipeline location, diameter, wall thickness, installation year, depth of cover with the incidents. These parameters are defined as ***background factor***. A background factor can be defined as one that seems to influence a causal factor under certain condition or value more than it does under other conditions, thus increasing the likelihood of failure due to that causal factor. They are generally inherent characteristics of the pipeline and transportation and do not have any direct influence on pipeline failure. For instance, pipelines with certain diameters seem to fail more than those with larger or smaller diameters and corrosion seems to be a leading causal factor behind such failures. Here, corrosion is considered as the causal factor with diameter of the pipeline being the background factor. It may be that the mechanism by which the background factors influence the causal factors is not yet perceived, and subsequently, why they influence the causal factors is not yet well understood. There is a lack of studies linking causal factors to the management system failures. Typically management, design, planning, organizational or operational failings are identified as root cause or underlying cause (Occupational Safety and Health Administration 2015, National Energy Board 2019), which is defined as "a fundamental, underlying, system-related reason why an incident occurred that identifies one or more correctable system failures" (Center for Chemical Process Safety 2019). In this article, the term ***underlying cause*** has been used subsequently. They represent the system's performance, and have a direct influence on a causal factor, but do not directly cause a pipeline failure. Similar studies in oil and gas production in offshore (Halim et al. 2018), onshore (Yu et al. 2017), and hazardous material transport (Quddus et al. 2018) incidents show a strong link between causal factors with underlying causes, as the negative events and undesirable conditions involve some of the active and latent failures that contributed to the incident.

Table 2.2 A summary of articles on causal analysis of pipeline incidents

Author (Year)	Causal factors	Background factors	Data sources**
(Andersen and Misund 1983)	Outside force/ third party damage, corrosion, mechanical failure, material and construction defects	Pipeline age, location, diameter, commodity transported	CONCAWE, US DOT
(Papadakis 1999)	Corrosion, external interference, construction/ material defect, other	Pipeline diameter, commodity transported, location	CONCAWE, EGIG, US DOT, VNIIGAS (Soviet Union)
(Bersani et al. 2010)	Corrosion, mechanical, third-party	*Hydrological, anthropogenic, technical factors	CONCAWE, US DOT
(Han and Weng 2011)	External interference, corrosion, material defect, operation error, ground movement	Flow rate, pressure, wall thickness, pipeline diameter, service life, depth of cover	US DOT GTG
(Cunha 2012)	Corrosion, material construction, natural causes, third-party action, others-unknown	Commodity transported, coating type, wall thickness, nominal diameter, population density, depth of cover	EGIG, CONCAWE, UKOPA, US DOT, Trans Petro, NEB
(Wang and Duncan 2014)	Corrosion, outside force, construction/ material defects	Pipeline age, location	US DOT GTG
(Siler-Evans et al. 2014)	Weather/ natural disaster, outside forces, operator error, material failure, corrosion, other		US DOT
(Lam et al. 2016)	Corrosion, material failure, excavation damage, other outside forces, natural forces	Location, pipeline material, pipeline age, diameter, corrosion prevention measure	US DOT GTG
(Ramírez-Camacho et al. 2017)	Third-party activity, corrosion, mechanical failure, operational/human error, natural hazards, equipment failure	Pipeline material, population density	MHIDAS

Author (Year)	Causal factors	Background factors	Data sources**
(Bubbico 2018)	Corrosion failure, equipment failure, excavation failure, incorrect operation, material failure of pipe or weld, natural force damage, other outside force damage, other incident causes	Commodity transported, pipe material, location, corrosion protection system	US DOT

*Hydrological (crossing of river, groundwater depth, zone of landslide, lithography, soil permeability); Anthropogenic (land use, population density, street crossing, railway crossing, sewage system, aqueduct crossing, electrical system, other utilities); Technical factors (operating pressure, diameter, wall thickness, burial depth, maximum operating pressure, specified minimum yield strength, year of construction, metal joint, number of internal and external imperfections, absence of metal in the imperfections.

**US DOT GTG – United Stated Department of Transportation, Gas Transmission and Gathering, NEB – Canada National Energy Board; EGIG – European Gas Pipeline Incident Data Group; CONCAWE – Conservation of Clean Air and Water in Europe; UKOPA – United Kingdom Onshore Pipeline Operators' Association; TransPetro – Petrobras Transporte S.A.; MHIDAS – Major Hazard Incident Data Service.

A summary of articles on causal analysis of pipeline incidents

The objective of the present study is set to understand how pipeline incident data was analyzed in the literature and identify the scope of improvement of the analysis. The study examines the frameworks used to classify the causal factors of the incident data from various databases and reports and found that the difference is minimal. Distribution of the causal factors and corresponding failure rates are compared for the datasets considered. Since one dataset allowed reporting of multiple causal factors for a single incident, it is studied further to understand its importance. Association between causal factors and a few selected background factors are investigated to identify the dependence of the factors. Influence of underlying factors on causal factors and their interdependence are also studied. Finally, the relative importance the different factors collected by the organizations and the limitations of the current analysis techniques are discussed. Associativity and causality of various types of factors and causes established to the pipeline incident are discussed. It is concluded that without a proper causal model, the understanding of the pipeline failure is partial and flawed.

2.2 Pipeline Incident Data

The current analysis investigates pipeline incident data from three data sources originating from three regions: US PHMSA, Canada National Energy Board (NEB), and European Gas Pipeline Incident Data Group (EGIG). US PHMSA maintains four separate incident databases for hazardous liquid (Banimostafa et al.), natural gas transmission and gathering (GTG), natural gas distribution (Wu et al.), and liquified natural gas (LNG). PHMSA pipeline infrastructure has 347,020 km (215,628 miles) of pipeline for crude oil, refined products, and natural gas liquids, 513,070 km (318,807 miles) of pipeline for gathering and transmitting natural gas, 3.5 million km (2.2 million miles) for distributing gas to homes, businesses, and other industrial sites, and a relatively smaller LNG pipeline network (Pipeline and Hazardous Materials Safety Administration 2019). Following the PHMSA incident reporting criteria, operators submit an incident report for each failure in their pipeline system. PHMSA authority then reviews the incident reports and updates the failure causes using a structured cause mapping. The structured cause mapping is a method by which the operator submission data is reorganized by PHMSA to fit into a structured map to bring consistency in the data that has been collected over the years. This dataset is known as *flagged data*, which contains failure causes as ‘mapped cause’ and ‘mapped sub-cause’. Mapped cause is a list of direct cause of failure (*e.g.*, corrosion, excavation damage) and mapped sub-cause is a list of more detail causes of mapped cause. For instance, sub-cause for corrosion which is a mapped cause includes internal

corrosion and external corrosion. Mapped cause and sub-causes are used in the current study. PHMSA started collecting the incident data in 1970 and the incident reporting system have gone through several changes since then. The current reporting system started in 2010 and contains 606 data fields (*e.g.*, date of incident, location, operator name, cause) for each incident. In the present study, incident records (flagged file) of HL database (3,755 incident records) and GTG database (1,157 incident records) for the period of 3/10/2010-5/31/2019 have been considered (Pipeline and Hazardous Materials Safety Administration 2019).

The Canadian pipeline incident database from NEB (National Energy Board) which covers 73,000 km of pipeline operation is also considered in the analysis. 1,297 incident records from 1/2/2008 – 4/4/2019 are used. It contains 102 data fields including apparent cause (what happened) and underlying cause (why it happened) (Canada Energy Regulator 2019).

EGIG is a cooperation of seventeen gas transmission operators in Europe (Gasconnect, Austria; Fluxys, Belgium; NET4GAS, Czech Republic; DGC, Denmark; Gasum, Finland; GRT Gaz, France; TIGF, France; Open Grid Europe, Germany; Gas Networks Ireland, Ireland; Snam Rete Gas, Italy; Gasunie, Netherlands/ Germany; REN Gasodutos S.A., Portugal; EUSTREAM, Slovak Republic; ENAGAS, S.A., Spain; Swedegas A.B., Sweden; SWISSGAS, Switzerland; National Grid, UK) (European Gas Pipeline Incident Data Group 2018). It has been gathering pipeline incident data collected since 1970 and publishing analysis reports based on the collected data. The 10th EGIG report, which is used in the current analysis, was published in 2018 (European Gas Pipeline Incident Data Group 2018). The report contains 208 incident records that were collected on 142,794 km of pipelines between 2007-2016.

Causal factors distribution and failure rates will be determined for failure data obtained from all three data sources, PHMSA, NEB and EGIG. The time period and span in which the data was collected are similar and comparable. A summary of data and definition used for the present analysis are given in Table 2.3.

Table 2.3 A summary of data and definition used for causal analysis of pipeline incidents

Data source	Year	Number of records	Mileage covered	Data used
PHMSA HL	2010-2019	3,755	347,020 km	Causal factors, Background factors
PHMSA GTG	2010-2010	1,157	513,070 km	Causal factors
NEB	2008-2019	1,297	73,000 Km	Causal factors, Underlying causes
EGIG	2007-2016	208	142,794 km	Causal factors,

Causal factor: a major unplanned, unintended contributor to an incident that if eliminated would have either prevented the occurrence of the incident, or reduced its severity or frequency

Background factors: have associations with pipeline failure and the likelihood of the failure due to the causal factor, however, because of their inherent characteristics they cannot be responsible for the failures

Underlying cause: a fundamental, underlying, system-related reason why an incident occurred that identifies one or more correctable system failures

2.3 Causal Factors in Incident Data

2.3.1 Comparison of causal factors

PHMSA classifies causal factors into 7 categories (*1. corrosion, 2. natural force damage, 3. excavation damage, 4. other outside force damage, 5. material/ weld/ equipment failure, 6. incorrect operation, and 7. all other causes*). With mapped sub-causes, cause-classification forms a causal-tree with more detailed information. For instance, the mapped cause "*Corrosion*" is sub-divided into "*Internal corrosion*" and "*External Corrosion*" as Sub-cause. Another level of information may also be available for some sub-causes such as type of corrosion (*e.g.*, general corrosion, localized pitting, or galvanic corrosion), information about cathodic protection.

NEB uses a similar classification for causal factors as PHMSA with 7 categories (*1. defect and deterioration, 2. corrosion and cracking, 3. equipment failure, 4. incorrect operation, 5. external interference, 6. natural force damage, and 7. other causes*). In addition to cause classification, NEB requires the reporting of 9 underlying causes of an incident (*1. engineering and planning, 2. maintenance, 3. inadequate procurement, 4. tools and equipment, 5. standards and procedures, 6. failure in communication, 7. inadequate supervision, 8. human factors, and 9. natural or environmental forces*). The one NEB feature that stands out is reporting of multiple causes for one incident suggesting that more than one cause can lead to a failure.

EGIG report uses 6 categories for the causal factors (*1. external interference, 2. corrosion, 3. construction defect/material failure, 4. hot tap made by error, 5. ground movement, and 6. other and unknown*). Each category contains additional information as well.

Table 2.4 compares the categories of causal factors responsible for pipeline failure in PHMSA, NEB and EGIG incident datasets. All three datasets have similar schemes for classification of causal factors. All three of them have *Corrosion, Natural Force Damage* (EGIG calls it *Ground Movement*), and *Other Cause* categories. PHMSA has two categories, namely, *Excavation Damage* and *Other Outside Force Damage* for *External Interference* as defined by both NEB and EGIG. On the other hand, both PHMSA and EGIG have a single category for *Material and Equipment Failure* as opposed to NEB's two separate categories for *Equipment Failure* and *Defect and Deterioration*. There are a few ambiguities or inconsistencies among the classification schemes, such as, it is inconclusive if the third-party damage is included in EGIG's *External Interference* category or *Incorrect Operation* is under *Hot Tapping made by Error*. Nevertheless, it is evident from the table that major cause categories from all three pipeline systems are based on similar principles and almost identical.

Table 2.4 Comparison of causal mapping of pipeline incident data

Pipeline Incident Causes and Sub-Causes		
PHMSA, USA	**NEB, Canada**	**EGIG, Europe**
Corrosion	**Corrosion and Cracking**	**Corrosion**
Internal corrosion External corrosion (*General corrosion, localized pitting, other*) (*Galvanic corrosion, atmospheric corrosion, stray current corrosion, microbiological corrosion, selective seam corrosion*)	**Damage or deterioration mechanism:** **Cracking** (*Fatigue, Corrosion fatigue, Stress corrosion cracking, Hydrogen induced cracking, Mechanical damage delayed cracking*) **Material loss** (*Internal material loss, External material loss, Poor condition of external coating, Disbondment, Holidays, Issue with impressed current*)	Appearance (*General, Pitting, Cracking*) In-line inspected (*yes, no, unknown*) Location (*Internal, External, Unknown*)
Excavation Damage	**External Interference**	**External Interference**
Operator/ contractor excavation damage, Previous damage due to excavation, Third party excavation damage	**Substandard conditions:** Congestion or restricted action; Defective tools (*Equipment or materials*); Fire and explosion hazards, Inadequate guards or barriers, Inadequate information or data, Inadequate instructions or procedures, Inadequate or improper protective equipment, Inadequate preparation or planning, Inadequate support or assistance, Inadequate warning system, Poor housekeeping or disorder, Presence of harmful materials **Weather related** (*Frozen components, High winds, Adverse weather, Heavy rains or floods, Temperature extremes, Wildland fire, Lightning*) **Damage or deterioration mechanism:** **External interference** (*Third party, Vandalism, Company contractor, Unknown*) **Geotechnical failure** (*Construction or undermining*)	Activity having caused the incident (*Digging, Piling, Ground works*) Equipment involved in the incident (*Anchor, Bulldozer, Excavator, Plough*) Installed protective measures (*Casing, Sleeves*)
Other Outside Force Damage		
Electrical arcing from other equipment/ facility, Fire/ explosion as primary cause, Fishing or maritime activity, Intentional damage, Maritime equipment or vessel adrift, Other outside force damage, Previous mechanical damage, Vehicle not engaged in excavation		
Incorrect Operation	**Incorrect Operation**	**Hot Tap Made by Error**
Damage by operator or operator's contractor, Incorrect equipment, Incorrect installation, Incorrect valve position, Other incorrect operation; Overfill/	**Damage or deterioration mechanism:** Other Causes (*Improper operation*) **Substandard acts:** Failure to check or monitor, Failure to communicate or coordinate, Failure to follow procedure or policy or practice, Failure to	

Pipeline Incident Causes and Sub-Causes		
PHMSA, USA	**NEB, Canada**	**EGIG, Europe**
overflow of tank/ vessel/ sump, Pipeline/ equipment over-pressured	identify hazard or risk, Failure to react or correct, Failure to secure, Failing to use PPE properly, Failure to warn, Horseplay, Improper loading, Improper placement, Improper position for task, Under influence of alcohol and/or other drugs, Using equipment improperly	
Material/Weld/Equip Failure	**Equipment Failure**	**Construction Defect/Material Failure**
Construction, installation or fabrication-related, Defective or loose tubing/fitting, Environmental cracking-related, Failure of equipment body, Malfunction of control/relief equipment, Manufacturing-related, Non-threaded connection failure, Other equipment failure, Pump or pump-related equipment, Threaded connection/coupling failure	**Damage or deterioration mechanism:** **Electrical power system failure** (*Electrical fault, Arc flash*) **Equipment** (*Valve seals or packing, Gasket/O-ring, Ancillary equipment*); Other Causes (*Control system malfunction*)	Defect details (*Hard spot, Lamination, Material, Field weld or unknown*) Pipeline component type (*Straight, Field bend, Factory bend*) Type of defect (C*onstruction or material*)
	Defect and Deterioration	
	Damage or deterioration mechanism: **Construction** (*Other defective welds, Defective other joint, Overbending, Defective pipe or component body, Wrinkle or buckle, Defective circumferential weld, Dent*) **Material or manufacturing** (*Defective circumferential weld, Defective pipe or component body, Defective longitudinal seam weld*) **Structural degradation** (*Corrosion fatigue, Other chemical degradation, Overheating, Weeping, Damage to reinforcement fibers*)	
Natural Force Damage	**Natural Force Damage**	**Ground Movement**
Earth movement, Heavy rains/floods, High winds Lightning, Other natural force damage, Temperature	**Damage or deterioration mechanism: Geotechnical failure** (*Scouring, Wash-out or erosion, Flotation, Soil subsidence/ slope movement, Frost heave, Landslide*) **Other causes** (*Unknown*)	Type of ground movement (*Dike break, Erosion, Flood, Landslide, Mining, River or unknown*)
All Other Causes	**Other Causes**	**Other and Unknown**
Miscellaneous, Unknown		Sub-causes out of category (*Design error, Lightning, Maintenance error*)

2.3.2 Distribution of causal factors

Table 2.5 presents the distribution of causal factors and the failure rate due to each factor from the PHMSA HL database (2010-2019), PHMSA GTG database (2010-2019), NEB database (2008-2019), and EGIG report (2007-2016) for time periods indicated in the parentheses. Pipeline failure rates are expressed in per 1,000 km-year for all causal factors *i.e.*, number of failures per 1,000 km per year for each causal factor. Total operating lengths (*i.e.*, pipeline mileage) as mentioned in section 2 are used to calculate the failure rate. Analysis follows the categories of causal factors described in the previous section and are compared in Table 2.3. Percentage distribution of causal factors are presented graphically in Figure 2.3. *Equipment Failure* for PHMSA HL, PHMSA GTG, and NEB data and *External Interference* for EGIG report are found to be the most frequently occurring causal factor. *Corrosion* appears to be the second most frequently occurring causal factor in PHMSA HL, GTG, and EGIG data and *External Interference* for NEB data. Failure rates for PHMSA HL and NEB data are higher than the PHMSA GTG or EGIG failure rates suggesting failure rate of hazardous liquid pipelines is higher than that of natural gas pipeline. It should be noted that PHMSA GTG and EGIG are gas transmission pipeline whereas PHMSA HL is hazardous liquid pipeline. NEB regulated pipelines include approximately two-third length of natural gas pipelines and the remaining one-third are liquid substance.

NEB dataset contains 21.4% incidents where multiple causal factors, termed as combination factors, have been reported as cause of failure and the next section will focus on such incidents.

Table 2.5 Number of pipeline incidents and their percentage distribution for different causal factors for PHMSA HL, PHMSA GTG, NEB, and EGIG datasets are presented. Number in the parenthesis indicates the total number of incident and percentage distribution is shown above that. All failure rates are converted to number of failures per 1,000 km-year.

Data source	US PHMSA HL (2010 – 2019)		US PHMSA GTG (2010 – 2019)		Canada NEB (2008 – 2019)		Europe EGIG (2007 – 2016)	
Causal factors	% (#) of incidents	Failure rate /1000 km-year	% (#) of incidents	Failure rate /1000 km-year	% (#) of incidents	Failure rate /1000 km-year	% of incidents	Failure rate /1000 km-year
Corrosion	20.1 (727)	0.227	19.1 (228)	0.048	11.0 (139)	0.163	25.0	0.037
External interference	5.8 (208)	0.065	18.5 (214)	0.045	17.1 (216)	0.253	**28.4**	0.043
Incorrect operation	14.1 (511)	0.159	5.6 (65)	0.014	10.8 (137)	0.160	3.9	0.006
Equipment failure	**45.2 (1635)**	0.509	**31.4 (363)**	0.077	**20.5 (259)**	0.303	17.8	0.027
Material failure	7.2 (260)	0.081	11.3 (131)	0.028	10.8 (137)	0.160		
Natural force damage	4.5 (161)	0.050	7.9 (92)	0.019	4.7 (60)	0.070	14.9	0.022
Others	3.2 (114)	0.036	5.5 (64)	0.013	3.6 (46)	0.054	10.1	-
Combination factors	-	-	-	-	21.4 (270)	0.316	-	-

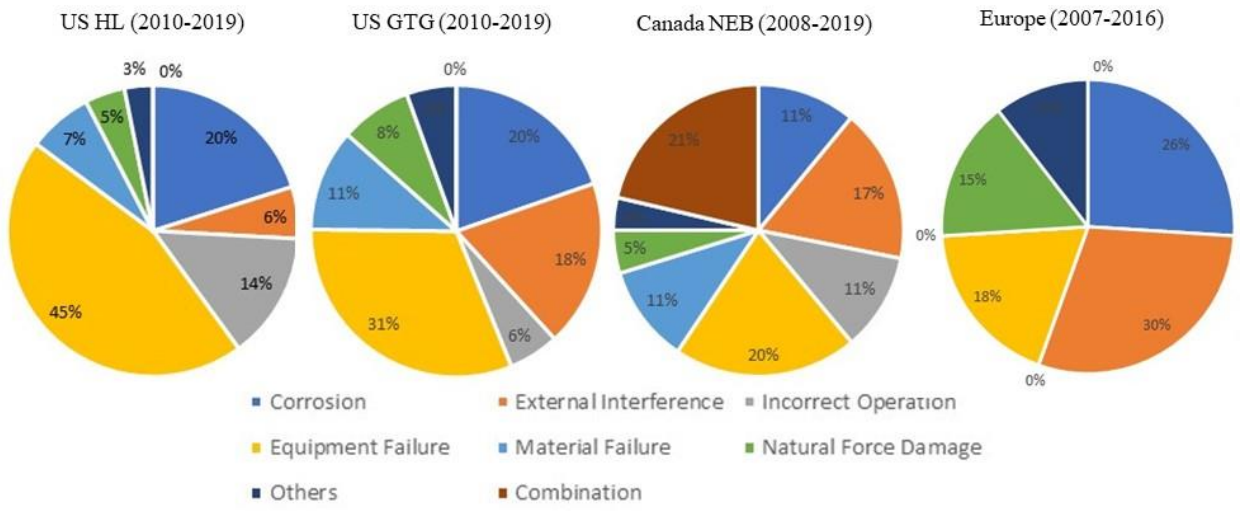

Figure 2.3 Distribution of causal factors for PHMSA HL, PHMSA GTG, NEB and EGIG incident data

2.3.3 Appearance of causal factors in combination

NEB database allows reporting of multiple causal factors for one incident, unlike PHMSA or EGIG datasets. Distribution of casual factors for NEB incident data (see Table 2.5 and Figure 2.3 in previous section) illustrates that multiple causal factors have been reported for 270 (21.4%) incidents. The analysis of these 270 incidents presented in Table 2.6 identifies the most frequently occurring combinations of causal factors and the associations amongst the causal factors. The second column of the table indicates the total number of incidents, where an individual causal factor appears (be it alone as a single cause, or in combination with other causes) and the third column indicates the total number of incidents where an individual causal factor appears only in combination with other causal factors. The subsequent columns represent the number of incidents where two causal factors occur as combination: two causes being the causal factors named in the corresponding row and column of the cell in which the number appears. For instance, *corrosion and cracking* appears in a total of 172 incidents as causal factor. It appears 139 times alone out of the 172 incidents and remaining 33 times in combination with other causal factors. With *defect and deterioration*, it appears 8 times, with *equipment failure* 5 times, with *external interference* 9 times and so on. It is worth noting at this point that in some incidents more than 2 causes, say 3, 4 or even 5 causes, were identified to occur in combination. But Table 2.6 shows association between two causal factors only. This means that if for a single incident 3 causes were identified (say *corrosion and cracking, defect and deterioration* and *equipment failure*) then Table 2.6 counted

this incident both in the cell that corresponded to *corrosion and cracking* with *defect and deterioration* as well as that which corresponded to *corrosion and cracking* with *equipment failure*. For this reason, the summation of the cause-combinations in the row corresponding to *corrosion and cracking* is 39, which is higher than the total number of incidents where causal factor *corrosion and cracking* appears as combination (33). This suggests that there are at most 6 incidents where more than two causes were reported. For this same reasoning, the summation of the values in the second column is greater than the actual number of incidents in the database.

The top three combinations of causal factors include *incomplete operation–external interference* (130), *external interference–equipment failure* (33), and *incorrect operation–equipment failure* (31). Some causal factors contribute significantly more in combination than when contributing alone: for example, 58% of failure due to *incorrect operation* (from Table 2.6 by dividing 192 by 329) and 48% of failure due to *external interference* are due to their combined effect (from Table 2.6 by dividing 203 by 419). For others, they alone contribute to a larger number of incidents than in combinations. Figure 2.4(a) shows the percentage distribution of the causal factors, based on third column (Total number of incidents where causal factor appears in combination) of Table 2.6, that appears as combination. This represents 21.4% of the total incidents in the database. This is another representation of *external interference* and *incorrect operation* being the top two contributors to combination causal factors. Figure 2.4(b) is based on column two of Table 2.6, which represents total number of occurrences of any causal factor alone or in combination in the entire database. It redistributes the causal factors in multiple cause incidents and merges it with incidents, where they also contribute alone. It shows a significant jump in contribution to incidents by *incorrect operation* and *external interference*.

Table 2.6 Multiple cause contributions for an incident (from NEB database)

Causal factor	Total number of incidents where the causal factor appears (alone or in combination)	Total number of incidents where causal factor appears in combination	Corrosion and cracking	Defect and deterioration	Equipment failure	External interference	Incorrect operation	Natural force damage	Other causes
Corrosion and cracking	172	33	139	8	5	9	14	3	0
Defect and deterioration	174	37		137	16	12	11	2	0
Equipment failure	333	74			259	33	31	3	1
External interference	419	203				216	130	17	5
Incorrect operation	329	192					137	4	4
Natural force damage	89	29						60	1
Other causes	54	8							46
Total	1570	576							

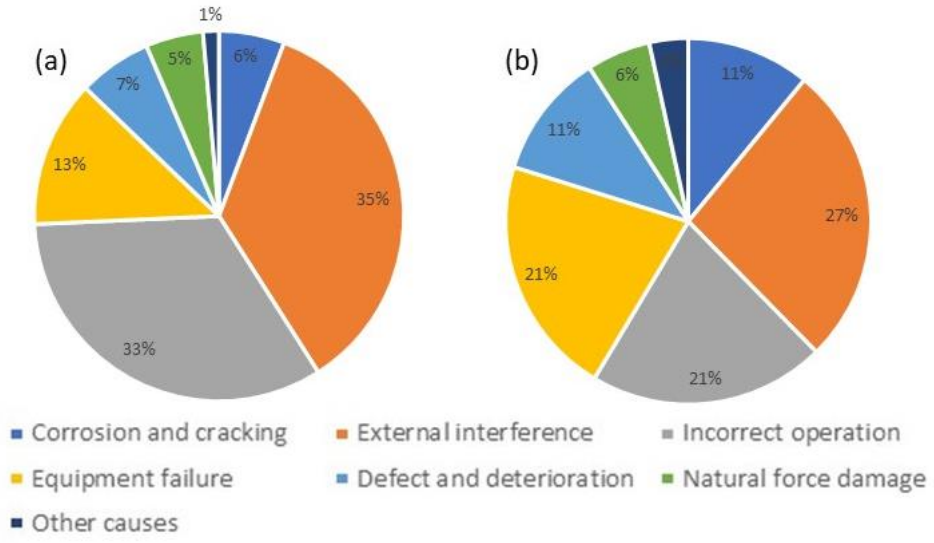

Figure 2.4 (a) Percentage distribution of causal factors involving multiple-cause failures (using the data on the third column of Table 2.6) (b) Modified distribution of cause contribution to pipeline incidents reported to NEB.

2.4 Background Factors in Incident Data

In the incident report form, PHMSA collects information about location, facility, operating conditions, and consequence as well as apparent cause of failure. The data includes condition of pipeline network at the time of the incident, and it does not change with the outcome of the incident. Some of the parameters can be associated with the causal factors shown in Table 2.2, such as, commodity transported, year of installation *i.e.*, age of the pipeline network, pipeline diameter, pipeline thickness, pipeline material, operating pressure, depth of cover, type of operation, population density because of their specific characteristics. However, these are not the causes of failure. For instance, consider three such parameters: commodity transported, pipe diameter and year of installation. These are presented in Table 2.7. Each of the parameters has several categories. For example, commodity transported includes crude oil, non-HVL (highly volatile liquid), HVL (highly volatile liquid), CO2 and biofuel. For each of the commodities, number of failures for each causal factor are presented. In a similar way data for pipe diameter and installation year are presented in Table 2.7. The data shows that for certain values of the parameters, the number of causal factors (and hence the number of incidents) are significantly greater than for other values. For example, the total number of incidents involving crude oil transportation is about 1.5 times than that of non-HVL, however, corrosion failure in crude oil pipelines is thrice that in non-HVL pipeline.

Table 2.7 Relationship with background factors with causal factors as obtained from US PHMSA HL data

		US PHMSA HL (2010 – 2019) Causal factors					
		Corrosion (752)	External Interference (214)	Incorrect Operation (536)	Equipment & Material Failure (1977)	Natural Force Damage (167)	All Other Causes (109)
Commodity Transported (Part B)	Crude oil	521	102	270	881	73	51
	Non-HVL	172	71	195	676	72	37
	HVL	51	40	62	383	22	18
	CO2	7	1	7	36	0	3
	Biofuel	1	0	2	1	0	0
Pipe Diameter in inch (Part H)	0 – 6	63	25	1	30	4	1
	6 – 12	242	93	14	82	15	14
	12 – 18	85	26	4	39	3	1
	18 – 24	60	7	5	21	1	3
	24 –	25	5	1	21	0	6
Installation Year (Part I)	Pre 1920	2	0	0	0	0	0
	1920-29	14	2	1	8	0	1
	1930-39	19	4	2	3	0	2
	1940-49	43	19	7	24	7	4
	1950-59	101	43	18	85	13	2
	1960-69	68	32	23	102	14	6
	1970-79	70	23	21	103	14	12
	1980-89	37	6	9	75	5	5
	1990-99	56	10	12	94	9	5
	2000-09	34	13	41	190	17	8
	2010-19	56	16	181	510	29	17

The background factors provide valuable insight regarding the causal factors, which otherwise cannot be extracted from overall data. Consider Table 2.8, 2.9, 2.10, representing association of three background factors (commodity transferred, pipe diameter, and installation year) with one causal factor: *corrosion*. The data from the third column of Table 2.7 regarding

corrosion failure is further processed for the three background factors. Table 2.8 presents the five categories of commodity transferred (crude oil, non-HVL, HVL, CO2, and biofuel), corresponding mileage, number of incidents, calculated percentage of incident, and calculated individual failure rate. The failure rate is calculated by normalizing (dividing) the number of incidents with the mileage and years of operation. Such analysis unravels the fact that the failure rate for crude oil transportation is almost double that of the average failure rate due to corrosion while the failure rate of HVL is significantly lower than the average failure rate due to corrosion (almost one-fifth). Similarly, Table 2.9 and Table 2.10 show the variations of failure rates with different categories of background factors. For pipe diameter, failure rates are lower for larger diameter pipes than smaller ones. A decreasing corrosion failure rate is observed for pipes with more recent installation dates. Similarly, more associations can be derived between other causal factors and background factors.

It appears that background factors can influence the causal factors but given the data it is not possible to identify the relationship between a causal factor with a given background factor. This is because a single background factor may not directly influence a causal factor; in fact, it is possible that multiple background factors may be at work behind a causal factor. However, the number of associations a single causal factor holds with a variety of background factors are large. In the data collected, for each causal factor, the background factors varied in each incident (different conditions of each parameter) and hence, drawing direct relations between all background factors with a particular causal factor, or understanding the dependency of causal factors on background factors, is not possible under the current circumstances.

Table 2.8 Association between commodity transferred and corrosion from US PHMSA HL data

Corrosion (Total incident: 752; failure rate: 0.227 failures/1000 km-year)					
Commodity	Mileage	# of incidents	% of incidents	Failure rate	% deviation from average
Crude oil	80750	521	69.3	0.427	88.1
Non-HVL	62711	172	22.9	0.181	20.3
HVL	70267	51	6.8	0.048	78.9
CO2	5206	7	0.9	0.089	60.8
Biofuel	15	1	0.1	4.408	1841.9
Total	218949	752		0.227	

Table 2.9 Association between pipe diameter and corrosion from US PHMSA HL data

Corrosion (Total incident: 752; failure rate: 0.227 failures/1000 km-year)					
Pipe diameter	Mileage	# of incidents	% of incidents	Failure rate	% deviation from average
0 – 6 in	34160	63	8.4	0.122	46.3
6 – 12 in	104641	242	32.2	0.153	32.6
12 – 18 in	29450	85	11.3	0.191	15.9
18 – 24 in	24625	60	8.0	0.161	29.1
24 – in	17930	25	3.3	0.092	59.5
Unknown	218949	277	36.8	0.084	63.0
Total	218949	752		0.227	

Table 2.10 Association between installation year (pipeline age) and corrosion from US PHMSA HL data

Corrosion (Total number of incidents: 752; average failure rate: 0.227 failure/1000 km-year)					
Installation Year (Part I)	Mileage	# of incidents	% of incidents	Failure rates (number of failures/1000 km-year)	% deviation from average
Pre 1920	479	2	0.3	0.276	21.6
1920-29	1907	14	1.9	0.485	113.7
1930-39	5051	19	2.5	0.249	9.7
1940-49	14821	43	5.7	0.192	15.4
1950-59	33783	101	13.4	0.198	12.8
1960-69	34080	68	9.0	0.132	41.9
1970-79	29930	70	9.3	0.155	31.7
1980-89	17609	37	4.9	0.139	38.8
1990-99	18687	56	7.4	0.198	12.8
2000-09	16956	34	4.5	0.133	41.4
2010-19	36896	56	7.4	0.100	55.9
Unknown	218949	252	33.5	0.076	66.5
Total	218949	752	100.0	0.227	

2.5 Underlying Causes in NEB Incident Data

2.5.1 Distribution of underlying causes

Unlike PHMSA and EGIG, NEB collects underlying causes ("*why it happened*") in addition to causal factors ("*what happened*"). Some organizational or management system elements are identified as underlying causes that may contribute to any causal factor. Distribution of underlying causes on how they affect the incidents are plotted in Figure 2.5. The NEB allows identification of multiple causal factors in an incident. It also allows multiple underlying causes to be identified for each incident but does not relate the causal factors with the identified underlying causes. *Maintenance* (35%), *Engineering and Planning* (19%), *Human Factors* (12%) and *Standards and Procedures* (11%) are identified as the top four dominant underlying causes. These underlying causes are defined by NEB as follows (National Energy Board 2019):

Engineering and Planning: failures of assessment, inadequate planning or monitoring, inadequate specifications or design criteria, lack of evaluation of change, or implementation of controls

Maintenance: inadequate preventive maintenance or repairs, failure to maintain excessive wear and tear

Inadequate Procurement: failures in the purchasing, handling, transport and storage of required materials

Tools and Equipment: improper use or inadequate tools and equipment

Standards and Procedures: inadequate development, communication, use, maintenance or monitoring of standards and procedures

Failure in Communication: loss of communication with automatic devices, equipment or people

Inadequate Supervision: lack of oversight of a contractor or employee during construction or maintenance activities

Human Factors: individual conduct or capability, or physical and psychological factors, and

Natural or Environmental Forces: external natural or environmental conditions.

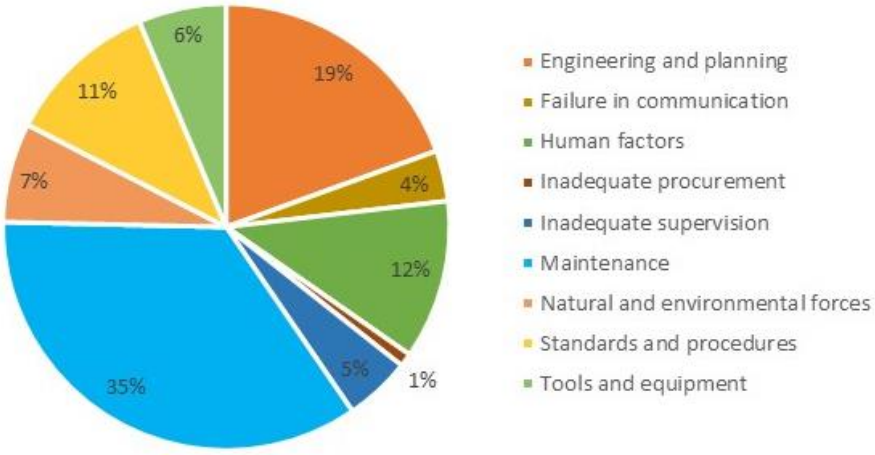

Figure 2.5 Percentage distribution of underlying causes of pipeline failures data from NEB

2.5.2 Relation of underlying causes and causal factors

Associations between the causal factors and underlying causes of NEB pipeline failure data are presented in Table 2.11. The number in each cell represents the number of incidents where a causal factor and an underlying cause appeared together. In other words, it expresses what underlying cause acted behind the failure of an incident for each causal factor. For instance, *maintenance* is reported as an underlying cause in 102 incidents out of 172 incidents where *corrosion and cracking* was reported as a causal factor. Thus, poor maintenance is responsible for almost 60% of corrosion related incidents. All associations with 100 or more incidents between a causal factor and an underlying cause are highlighted with amber and associations with more than 50 but less than 100 incidents with yellow (see Table 2.7). It is evident from the analysis that *maintenance* is the single most important underlying cause affecting all major causal factors. *Engineering and planning* and *human factors* are other two significant underlying causes. Another interesting observation is that *corrosion and cracking*, *defect and deterioration*, and *equipment failure* are primarily affected by a single underlying cause *maintenance*. Similar remark can be made for *natural force damage* and *natural or environmental forces*. However, *external interference* and *incorrect operation* are influenced by multitude of underlying causes suggesting management of these causal factors will be much more challenging than the others because of the possible interplay among the underlying causes.

Table 2.11 Relationship with causal factors and underlying causes as obtained from NEB data

		Total number of incidents where the causal factor appears	Underlying causes								
			Engineering and planning	Failure in communication	Human factors	Inadequate procurement	Inadequate supervision	Maintenance	Natural or environmental forces	Standards and procedures	Tools and equipment
Causal factors	Corrosion and cracking	172	63	4	6	1	2	102	6	17	10
	Defect and deterioration	174	41	5	13	4	5	124	3	32	6
	Equipment failure	333	62	5	29	5	14	222	15	30	22
	External interference	419	134	36	93	8	43	111	55	76	66
	Incorrect operation	329	64	43	131	9	65	91	8	81	54
	Natural force damage	89	27	0	0	0	0	18	67	2	0
	Other causes	54	10	0	4	0	0	37	2	0	4

2.5.3 Association among underlying causes

To understand the interplay among the underlying causes, the associations of the underlying causes as evident from NEB data are presented in Table 2.12. This is like Table 2.6. Analysis determines the number of appearances where the underlying causes act alone (diagonals) in the background to an incident and where they work in combination (non-diagonals) with others. *Maintenance* (431 alone out of 581 total incidents: 74%) and *natural or environmental forces* (99 alone out 120 total incidents: 83%) appears to be a type of causes that work alone most of the time. On the contrary, *engineering and planning* (205 alone out of 324 total incidents: 63%), *human factors* (78 alone out or 192 total incidents: 41%), *standards and procedures* (83 alone out of 184 total incidents:

45%), and *tools and equipment* (46 alone out of 106 total incidents: 43%) appear to act together with other underlying causes. Significant and moderate combinations are highlighted in amber and yellow, respectively.

Table 2.12 Dependencies of underlying cause contributions as obtained from NEB data

	Number of incident where the underlying cause	Engineering and planning	Failure in communication	Human factors	Inadequate procurement	Inadequate supervision	Maintenance	Natural or environmental forces	Standards and procedures	Tools and equipment
Engineering and planning	324	205	19	31	6	13	61	11	23	15
Failure in communication	61		14	19	0	8	8	0	9	7
Human factors	192			78	4	23	20	1	24	20
Inadequate procurement	16				2	1	1	1	1	0
Inadequate supervision	80					19	2	0	1	2
Maintenance	581						431	6	34	12
Natural or environmental forces	120							99	0	2
Standards and procedures	184								83	2
Tools and equipment	106									46

2.6 Pipeline Incident Investigation Reports

2.6.1 Available incident investigation reports

In an excel spreadsheet, there were records of 109 incident investigation reports, as shown in Appendix A1. Out of these 109 reports, 64 of them are of type “hazardous liquid”, 40 of them “gas gathering and transmission”, and the remaining 5 of them are of type “natural gas”. There were 81 incident investigation reports downloaded from PHMSA website. 32 of the reports where incidents occurred before 2010 and remaining 77 occurred after 2010. 26 of them involved corrosion, 43 of them equipment/ material failure, 9 of them excavation, 15 of them incorrect operation/ operator error, 5 of them natural force damage and a few for other reasons. Out of these 109 reports

mentioned in the excel spreadsheet, 81 of them could be successfully downloaded as pdf files. To make the analysis consistent with the incident records, incident reports prepared after 2010 involving hazardous liquid have been selected. 64 of such reports were identified. However, only 41 of the reports or in other words pdf files were readable. A summary of these reports was given in Appendix A2.

The summary showed that incident investigation reports are varied by possible root causes and the investigation reports are not very lengthy suggesting not in-depth. Mostly, investigations were conducted to examine one specific topic and the reports were made accordingly.

2.6.1 General Structure of the Incident Investigation Reports

Although the purpose and the techniques used for the incident investigations varied, there is a general common structure identified in the reports. Some common features identified are:

- General Introductory Form
- Executive Summary
- System Details
- Events Leading Up to the Failure
- Emergency Response
- Summary of Return-to-Service (Also referred to as return to service, or summary of restart plan and return-to-service, or preliminary safety measures)
- Investigation Details
- Findings and Contributing Factors
- Appendix

A few less used headings include

- Pipe Specifications
- Analysis
- Conclusion

We have highlighted before that it would be impossible for someone to study all incident investigation reports manually and make an expert and reasonable judgement of the root causes or contributing factors. From this perspective, efforts have been made to identify the sections and sub-sections that can be studied or examined using a computer. For instance, the photographs or images or plots are a great way of conveying the findings to the reader. Sometimes information

and data gathered are presented in tabular fashion very efficiently and effectively. However, when viewed from the context of usefulness to a computer code, such tools (images, plots, or tables) are identified as less effective. Rather long detail conversational aspect, typically avoided in brief reports, are found interesting and useful. Some sections that were found useful are:

- Executive Summary, Conclusion
    - Contain a basic overview of the most useful information pertaining to the pipeline failure
- Events Leading Up to the Failure, Emergency Response, Summary of Return-to-Service
    - Contain information about what happened before, during and after the pipeline failure
- Investigation Details, Findings and Contributing Factors
    - Contain useful information about the investigation into the pipeline failure and the pipeline failure itself
- Analysis
    - Contains information about post-incident analysis of pipeline materials
    - Could be very useful for cases where corrosion has occurred as this section will provide in-depth detail about material damage, etc.

2.7 Comparison of Descriptions of Incident Data and Findings of Incident Investigation Reports

In this study, two different datasets were used to identify the causal factors behind the pipeline incidents. However, discrepancies may exist between the incident records (reports made right after the incidents) and findings from the incident investigation reports. To understand the anomalies or differences, 13 corrosion-related incidents in two data sources and 7 randomly selected not corrosion related incidents were identified and studied, as shown in Table 2.13

Table 2.13 List of incident investigation reports that were considered for the analysis

Report #	Investigation Report File Name	Map Cause
20100045	Whitecap OCS HL 2010-03-25 508	Other Outside Force Damage
20100054	Sunoco RM HL PA 2010-03-25	Material/ Weld/ Equipment Failure
20100146	Chevron HL UT 2010-06-11	Other Outside Force Damage
20100147	Suncor HL WY 2010-6-14	Incorrect Operation
20100163	Dixie_HL_GA_20100705	Excavation Damage
20100179	Magellan Ammonia HL NE 2010-07-23	Material/ Weld/ Equipment Failure
20100317	Chevron HL UT 2010-12-01	Incorrect Operation
20100014	Mid-Valley Pipeline HL TX 2010-03-01	Corrosion
20100042	SFPP HL CA 2010-03-16	Corrosion
20100287	Shell_Pipeline_2010-11-16	Corrosion
20110080	133500_Sunoco_2011-2-8	Corrosion
20110120	Buckeye_HL_PA_2011-03-20	Corrosion
20120141	Enterprise_West_Tank_Farm_Cushing_OK_2012-4-8	Corrosion
20120232	Buckeye HL PA 2012-07-13	Corrosion
20120366	Magellan 2012-11-25	Corrosion
20130130	Lion_Oil_Magnolia_Tank_2013-03-09	Corrosion
20130208	Enbridge 2013-05-17	Corrosion
20140333	Buckeye HL NJ 2014-8-20	Corrosion
20150224	Plains_Pipeline_LP_2015-5-19	Corrosion
20150464	Enterprise_2015_12_1	Corrosion

The records below summarized findings of the comparisons between incident records obtained from the narratives and the select sections of the incident investigation reports. Sections that were selected from the incident investigation reports are

Section List = ['Contributing Factors', 'Contributory Causes', 'Findings', 'Findings & Contributing Factors', 'Findings and Contributing Factors', 'Investigation Findings & Contributing Factors', 'Investigation Findings and Conclusions', 'Investigation Findings and Contributing Factors']

This comparison has been made manually to understand the reports better and devise a plan for the natural language processing technique.

Table 2.14 List of incidents that were compared between incident reporting database and incident investigation reports

Parameter	Report #	Description
Level of causes	20100287	Incident record narrative mentioned both apparent cause "external corrosion" and deeper cause “operational pressure fluctuating" Incident record narratives contain findings after investigation was done thus had same level of cause information that was mentioned in investigation report.
	20120141	Incident record narrative only mentioned "internal corrosion" Investigation report mentioned deeper cause which was related to maintenance
	20120232	Both Incident record narrative and investigation report mentioned corrosion was due to low cycle fatigue cracking While investigation report provided more insights saying placement of unused leak detection tubes near the failure location might contribute to low cycle fatigue
	20140333	Incident record narrative mentioned apparent cause “internal corrosion” and mentioned it was likely caused by MIC Investigation report stated multiple factors leading to MIC
	20120366	Both Incident record narrative and investigation report mentioned crevices and atmospheric corrosion. Report also mentioned deeper causes which are related to pipe design and pipe location which did not allow convenient atmospheric corrosion inspection.
	20130208	Incident record narrative did not mention apparent cause - internal corrosion or bacteria in the narrative. Investigation report does not have a section that belongs to the section list. Apparent causes and organizational causes such as no inhibition and maintenance issue were mentioned in “Conclusions”
	20150224	Investigation report detailed many issues with insufficient detection systems, ineffective protections against external corrosion, and a lack of timely response. These issues were not discussed in the incident record narrative
Data structure	20100317	Deeper cause due to lack of valve winterization program was not mentioned in incident record narratives, but was mentioned in the field “Operation Details”
	20130208	Investigation report does not have a section that belongs to the section list. Apparent causes and organizational causes such as no inhibition and maintenance issue were mentioned in “Conclusions”
	20130130	Investigation report does not have a section that belongs to the section list. Only one sentence mentioned the cause - deposit corrosion in the section of "conclusions"

Parameter	Report #	Description
	20130208	"Findings and contributing factors" section does not mention causes leading to failure, instead, only talks about inappropriate detection and responses. Another section "Conclusions" mentions the apparent cause “internal corrosion” and organizational causes leading to inappropriate responses
	20150464	Investigation report does not have a section that belongs to the section list. Only one sentence mentioned internal corrosion in the section "Conclusions"
Different terminologies	20150464	Incident record narratives did not mention “internal corrosion”, instead, it said "the pinhole was caused by carbon dioxide attack of the pipe"
Facts not contributing to incidents	20100042	Incident record narrative contains a long paragraph about details of responses “Findings and contributing factors” section in the investigation report contains facts such as "no indications of cracks or corrosion were found..."
	20100054	Incident record narrative mentioned "no corrosion of the flange faces or other mechanical damage was observed"
	20100287	“Findings and contributing factors” section in the investigation report contains the facts such as "no manufacturing defect", “did not determine whether MIC contribute to failure"
	20120141	Incident record narratives mentioned "inhibitor has been added"
	20140333	Investigation report mentioned "no evidence of external corrosion" Incident record narratives mentioned “A six day investigation into the incident involved excavation, isolation, and pressure testing of the 12 inch bayway line shipper manifold area of the station”. Excavation here was not causes but was one of investigation activities Incident record narrative mentioned "there was no pressure indicated or recorded that would show the failure was caused by a pressure above the mop."
Varying content	20108080 20100287 20110120 20140333 20130130	Extracting knowledge about event chains leading to failures can be difficult. Most narratives do not provide such information but mentioned a lot about how responses were taken.

In summary, most of the incident record narratives mention apparent causes clearly. Compared to the level of causes that are covered in incident record narratives and investigation report, investigation reports provided causes other than apparent causes more often. Incident record

narratives formats are consistent, and the inputs may not be consistent. causes could be shown in multiple fields. On the other hand, investigation report formats were not consistent. Most reports have a section in the section list to discuss about contributing factors, but some reports do not have such section. Inconsistent terminologies may bring difficulties in interpreting text mining results. Incident record narratives and the interesting sections in investigation reports could have facts which are not contributing to incidents. Such information could be misinterpreted as causes. Cautions are needed.

3.0 EXTRACTION OF NECESSARY INFORMATION FROM INCIDENT DATA AND INCIDENT INVESTIGATION REPORTS USING NLP

3.1 Objective

To develop an ANN model, data from past incidents need to be gathered. All incidents must be investigated to determine what went wrong and data must be recorded in a consistent manner. Different root causes analysis or incident investigation techniques have been adopted in the past to identify different causes and investigation reports expressed the causes in a variety of ways. Root cause or failure analysis had different fields of origin (such as safety-based root cause failure analysis, production-based root cause failure analysis) and what may appear as root cause may provoke further questioning to determine deeper hidden causes in another. Thus, for the selection of inputs, it would be essential to determine a reference that defines what will be termed as root cause. Using taxonomy so that similar terms are used to refer to related root causes can help tackle this issue. At the same time, identified root causes may be present in a way that cannot be used for measurement. This problem is increased when human and organizational factors are identified as root cause. For example, an investigation may find lack of maintenance as a root cause. This does not provide any measured value that can be used as an input. However, if the percentage of deferred maintenance and the length of deferred maintenance were measured for the system at the time of incident, then it would provide a quantitative assessment of the condition the pipeline system was in at the time of the incident. Thus, investigation findings will have to be recorded in a manner that will enable information related to the root cause to be expressed in terms of quantifiable indicators. For a similar reason, the output for the investigations will have to be expressed in terms of severity levels.

Thus, the first challenge would be to identify a methodology to build a set of cue words or taxonomy so that root causes analysis conducted for different incidents identify similar causes using similar terms and these causes will have to be identified in terms of measurable deviations/indicators so that they can be compared with deviations existing in a system to understand if the system is reaching an unsafe state. It will produce a consistency among the reports of root cause analysis to enable extraction of information from those reports to build a learning model and then compare them with the current condition of the system to predict failure. If a set of cue words are developed to produce all reports, extraction of information using automated systems based on text mining or data mining can be used.

3.2 Current Approach Using Natural Language Processing (NLP)

Incidents records collected by several federal agencies, such as the Pipeline and Hazardous Materials Safety Administration (PHMSA), Occupational Safety and Health Administration (OSHA), or Bureau of Safety and Environmental Enforcement (BSEE) form large databases, and can be a great learning resource if properly utilized (Yu et al. 2017, Halim et al. 2018, Quddus et al. 2018, Halim et al. 2020). The key to learning from the past incident records is to identify the underlying causes of the incidents (Halim et al. 2018, Halim et al. 2020). In the case of PHMSA, incident records are collected from operator submission and then post-processed to a more structured format for the ease of analyzing pipeline incidents through years. The post-processed incident records include a cause category and a short description for each incident, associated with much other information (Pipeline and Hazardous Materials Safety Administration 2019). There are seven cause categories pre-defined to categorize direct causes of the incidents that are reported by operator, including corrosion, equipment failure, material failure, natural force damage, excavation damage, incorrect operation, and others. Taken advantages of the structured information in the PHMSA database, extensive studies have been conducted on the direct causes and contributions from relevant other factors on pipeline failure (Bersani et al. 2010, Cunha 2012, Bubbico 2018, Halim et al. 2020). However, none has focused on the underlying causes and contributory factors such as organizational, managerial or personnel issues regarding the failure since such information is not always reported by operators and it only can be addressed in free-text incident descriptions if applicable (Pipeline and Hazardous Materials Safety Administration 2019). Thus, incident descriptions can be a great resource to identify underlying causes and contributing factors of pipeline incidents. Searching through thousands of such descriptions is not only tedious but almost humanely impossible. Given the usefulness of identify underlying causes (Pyun et al. 2020, Zhang et al. 2020) and contributory factors (Adedigba et al. 2016, Naghavi-Konjin et al. 2020), it would be interesting and worthwhile to explore the capability of natural language processing (NLP) as an option to automatically extract valuable knowledge from pipeline incident descriptions.

NLP is primarily concerned with programming computers to process, understand, interpret, and manipulate human language (Manning and Schütze 1999, Manning et al. 2014). NLP can be used in a variety of different tasks, fields and industries for sentiment analysis, text classification, question answering, *etc*. Text mining, also known as text data mining, is referred to the process of

transforming text data to numeric data that can be then analyzed by data mining algorithms (Miner et al. 2012). Applications of NLP or text mining techniques in the field of process safety have been reported to automate content analysis of vast amount of incident text data(Chokor et al. 2016, Tanguy et al. 2016, Tixier et al. 2016, Goh and Ubeynarayana 2017, Nakata 2017, Syeda et al. 2017, Verma and Maiti 2018, Zhang et al. 2019, Single et al. 2020). There are mainly two approaches to develop an automated content analysis system (Allahyari et al. 2017): NLP with hand-coded rules (Tixier et al. 2016, Nakata 2017, Verma and Maiti 2018, Single et al. 2020), and (2) NLP with machine learning algorithms (Tulechki 2015, Chokor et al. 2016, Tanguy et al. 2016, Goh and Ubeynarayana 2017, Syeda et al. 2017, Zhang et al. 2019). The first approach is to develop an NLP system based on pre-defined causality and/or dictionaries of key words by human experts. Tixier et al. (2016) utilized this approach to extract precursors and outcomes from construction injury reports. By manually pre-defining a list of words that mean causes and effects based on the attribute-based framework proposed by Esmaeili and Hallowell (2012), the developed NLP algorithm is able to scan incident reports, detect words that match the pre-defined tokens, and generate a summary tabulating all the key words for each report. Even though this method exhibits satisfying accuracy, it requires intensive labor to develop domain-specific dictionaries and results in loss of information as text data is manually reduced to limited tokens (Robinson et al. 2015). Nakata (2017) proposed a text-mining method to construct the flows of events based on aviation incident reports by extracting meaningful words (i.e., noun, proper noun, verb, and adjective) from a bag-of-word of neighboring two sentences. The underlying assumption is that one verb is indicative of one event. However, order of words was ignored in the study as reported which otherwise may have provided more information about causality.

To improve the autonomous capability of the content analysis system, the second approach of combining NLP with machine learning algorithms, is employed. Researchers investigated the strength of support vector machine (SVM) on classification of aviation incident reports (Tanguy et al. 2016). Although classification of some events can achieve the accuracy above 95%, the results are not consistently satisfying. Later on, six supervised machine learning algorithms including SVM, K-nearest neighbor (KNN), decision tree, logistic regression, random forest, and Naïve Bayesian are evaluated on classification of 1,000 construction accident narratives, finding that SVM produces the best results with precision ranged from 0.5 to 1 (Goh and Ubeynarayana 2017). An even better classification performance is achieved by an ensemble model consisting of

five classifiers including 6 algorithms except random forest. The weight of each base classifier in the ensemble model is optimized with cross-entropy loss as the objective function by sequential quadratic programming (Zhang et al. 2019). While these works illustrate how supervised learning techniques can be used in classifying documents, it is only applicable when pre-defined categories exist in the dataset. To explore the strength of unsupervised learning, Chokor et al. (2016) employed NLP with K-means clustering to categorize incidents based on the type of incident from the incident description. The limitations of this study include the sample size and specificity of the geographical area used. Use of topic modeling with NLP techniques to infer the latent structure of entities and build a causation model was found promising with a showcase of 6 railway incident reports (Syeda et al. 2017). There is a need to gain better understanding into the strength of NLP and unsupervised learning techniques on analyzing the incident text data and inferring causal relationships.

The current work focuses on using NLP and text mining techniques to extract contributing factors and latent causality of pipeline incidents. Instead of classifying documents, both K-means clustering, and co-occurrence network approach are employed to examine 3,587 incident narratives collected from PHMSA incident database to generate clusters of words that are likely to be contributory factors and form causal dependency. Techniques of dimensionality reduction, including principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), are evaluated regarding the clustering application. Even though PCA has been extensively used in process safety, such as identifying the inherently benign path of chemical synthesis (Srinivasan and Nhan 2008, Banimostafa et al. 2012), early detection of the process fault (Pyun et al. 2020), and human factors assessment (Omidi et al. 2018), this work demonstrates the limitation of PCA and the advantage of t-SNE when dealing with highly nonlinear dataset.

The remaining of this paper is organized as follows. In Section 2, the procedure of incident data collection and preprocessing, methods of text analytics, and the overall workflow of analysis are set out. Section 3 presents the results: causation model developed based on K-means clustering and co-occurrence network analysis. Section 4 concludes the work done.

When pipeline operators report pipeline incidents to PHMSA, the major cause of incident is selected from a pre-defined list in the reporting form. A supplemental “comment” section in the reporting form allows the operators to provide a narrative description on the incident from where any additional contributing factors can be extracted to form a more comprehensive causation

model of the incident. Due to the unstructured nature of the "comment" section, NLP with K-means clustering, and text mining techniques of co-occurrence network are applied to mine the hidden knowledge. Before employing the techniques, data collection and preprocessing steps are required.

3.3 Methodology Involving NLP Extracting Necessary Information

As mentioned previously, the source of data in the present work is the incident narratives from the "comment" section of each incident record. Certain preprocessing steps are used to transform the text data into a more digestible form for machine learning and text mining algorithms. It should be noted that even though preprocessing is considered common for NLP tasks, the operations and the sequence of steps are not always agreed upon due to the widely varying nature of task. In the current work, the Natural Language Toolkit (NLTK) package is applied for the following preprocessing steps (Loper and Bird 2002):

Tokenization: In tokenization, the text is split into tokens, which are essentially the individual words that make up a string of text. For instance, in the string "a mechanical seal on the pump failed" there are 7 total tokens.

Noise removal: This step is to remove tokens which barely add value to and/or even interfere with text analysis. The removed tokens include punctuations, whitespace, number, dates, and *stopwords*. Numbers and dates are excluded as these tokens are hardly relevant to contributing factors and causality. Thus, removed are *stopwords* which are referred to as extremely common words yet providing little value to analysis of the content, such as "an" and "the". The *stopwords* list used in the present work is from NLTK corpus.

Lemmatization: Lemmatization is the process of reducing words down to their base forms, or lemmas. For instance, words like *am*, *are*, and *is* can all be reduced to *be*. Lemmatization considers the context of a word as it is used in a sentence. For instance, the word *pump* can be a verb or a noun depending on the context of its usage. Attention to context is what differentiates lemmatization from stemming. Stemming can be used as a faster alternative to lemmatization. Though, what stemming gains in speed, it sacrifices in linguistic accuracy (Toman et al. 2006). Words like *pump* for instance, might only have one stem, regardless of whether the word is used as a noun or verb. In general, the current project values accuracy over speed, so lemmatization of words gets the priority over stemming instead.

Filtering: Besides noise removal, an additional filter is set as the last step of preprocessing to improve the performance of NLP and text mining techniques. There are meaningful words but too general to create insights. For instance, the word "pipeline" does not belong to any published list of *stopwords*, but in this work, it provides no additional information because it has been known that the text data are collected from a pipeline database. Moreover, the existence of "pipeline" may interfere the construction of causal model as it can possibly appear in the results of word clusters created by text mining algorithms. The list of filtering words is determined by a trial-and-error method based on the word clusters. A detailed discussion on filtering step is presented in Section 3.3.3.

3.3.1 Text to features

After preprocessing, narratives of each incident record are converted to clean and normalized tokens. To apply machine learning and text mining techniques, these tokens are further transformed into numeric features representing the text dataset. Two methods of feature extraction are presented in this section.

Term frequency-inverse document frequency (TF-IDF)

A weighing scheme of term frequency-inverse document frequency (TF-IDF) was first proposed by Jones (1972) to evaluate the importance of words in a collection of documents and has gained popularity in the field of NLP and text mining (Salton and Buckley 1988, Ramos 2003). TF-IDF calculates the relative frequency of each unique word in a specific document via inverse proportion of the documents containing that word, thus producing a term-document matrix of TF-IDF scores. Under this method, higher weights are assigned to the words that are not commonly observed across the dataset but obtain high frequency in a few documents. The following formula is applied to calculate TF-IDF value for a term *t* in a document *d* in the term-document matrix:

$$TF - IDF(t,d) = tf(t,d) * idf(t) = \ tf(t,d) * \log\left(\frac{N}{df(t)} + 1\right)$$

where *tf(t, d)* is the frequency of term (or word) *t* in document (or incident record) *d*, *N* is the total number of documents in the dataset, *df(t)* is the number of documents containing the term *t* in the document set, namely the document frequency. To account for the scenario where a term appears in every document, resulting in zero *idf(t)*, one is added inside the logarithmic term for *idf(t)* calculation. As the equation above generates the TF-IDF score for every term *t* in every document *j*, the TF-IDF matrix is produced with dimension of *N*M*, where *M* is the total number of unique

words in the dataset. The matrix then becomes the input dataset to be analyzed by K-means clustering algorithm.

Co-occurrence matrix

Unlike the term-document matrix of TF-IDF weights, co-occurrence matrix defines entities both in rows and columns as a unique word present in the text dataset, thus a word-word matrix to evaluate linkage between words. A co-occurrence of two words is determined when they both appear in the same document with the distance less than a certain window size (Veling and Van Der Weerd 1999). Thus, word co-occurrence works irrespective of appearance frequency. By scanning through the pre-processed text dataset, a co-occurrence matrix is built with dimension of *M*M*, where *M* is the total number of unique words present in the dataset. Under this treatment, each unique word is vectorized in terms of its co-occurring frequency with other words, which plays a fundamental role to further develop co-occurrence networks.

3.3.2 Methods of text analytics

To extract contributing factors and latent causality of incidents, clusters (or networks) of words with strong connections need to be identified at first by NLP and text mining algorithms. Then construction of causation models from word clusters or networks can become feasible. The current work explores the strengths of two analytical methods, namely K-means clustering and co-occurrence networks.

K-means clustering

Clustering analysis has been one of the most important topics in unsupervised learning, and K-means clustering is the most used clustering technique (Chokor et al. 2016, Allahyari et al. 2017). To be noted, this study applies unsupervised learning even though there is a "cause" label for each incident record, because the current objective is to mine the hidden knowledge of narrative comments beyond a single cause classification assigned by PHMSA. When K-means clustering is applied to the 3587 incident narratives (or documents), the documents with similar statistical pattern of TF-IDF scores tend to fall into the same cluster, and thus the words with accumulated high TF-IDF scores can be used to induce the hidden association of events. In TF-IDF matrix, each incident narrative is represented by a vector containing *M* elements of TF-IDF score where *M* is the total number of unique words in the dataset. Thus, the distance of documents can be calculated based on the distance of vectors. With K-mean clustering, the vectorized documents in TF-IDF matrix are partitioned into K distinct clusters based on Euclidean distance to the centroid of a

cluster (Wagstaff et al. 2001). A greedy algorithm is utilized to minimize the objective function *J* formulated as the within-cluster sum-of-squares (WCSS):

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||m_j - x_i^{(j)}||^2 \quad (2)$$

where m_j is the centroid of the j^{th} cluster, *n* represents the total number of documents. Since it starts with an initial partition with K clusters and the objective function (WCSS) always decreases with an increase in the number of clusters K, thus it can only be minimized for a fixed number of clusters. Scikit-learn, a Python-based machine learning package (Pedregosa et al. 2011), is used in the present work to perform K-means clustering. The elbow method is employed to determine the number of clusters (K) by plotting the WCSS versus K. This method assumes the percentage of variance explained by the clustering algorithm as a function of the number of clusters and thus identifies the optimal K when the contribution of adding one more cluster becomes negligible (Bholowalia and Kumar 2014). The elbow plot of the K-means clustering in this study is displayed in Figure 3.1 and the "elbow" point is observed when the number of clusters is 5. This is considered a reasonable value as the PHMSA HL database has 7 pre-defined cause categories.

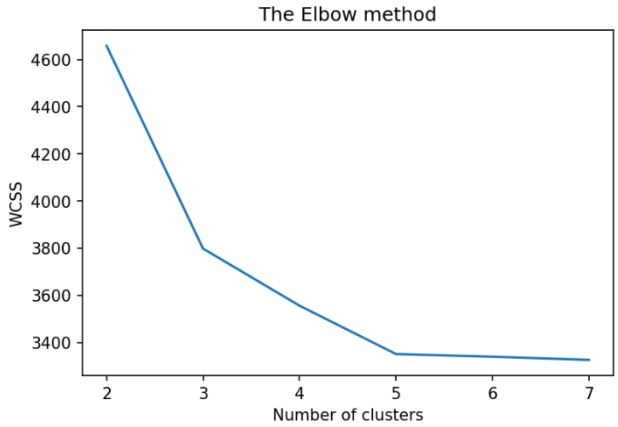

Figure 3.1 The elbow plot based on within-cluster sum-of-squares (WCSS) versus number of clusters

Topic Modeling

Topic modeling has been primarily employed to discover topics and latent relationships from a large set of text sources. By assuming that words occur or co-occur as a natural result of semantic pattern, any document in the context of topic modeling is viewed as a bag-of-word (a collection of terms without consideration of the appearing sequence) (Robinson 2019). As compared to unsupervised learning techniques (*e.g.*, K-means clustering) that relies on document-term matrix,

topic modeling splits the document-term matrix into the matrices of documents to a certain number of topics, and the matrices of individual topics to their constituent terms, resulting in identification of topics with weightings and the inferred constituent words. The differences between topic modeling approaches are on the statistical model applied to split the document-term matrix. Latent Dirichlet allocation (LDA), a Bayesian inferential statistical approach, as one of the most popular topic modeling approaches is utilized in the present work.

Co-occurrence network

Co-occurrence network is a graph of word interactions representing co-occurring patterns in the text data (Zhang et al. 2018) and has been widely used for many graph-based NLP applications, such as key object extraction (Mihalcea and Tarau 2004) and word sense discrimination (Ferret 2004). In the co-occurrence matrix, each unique word is represented by a vector containing elements of its co-occurrences with other words. Jaccard similarity coefficients, a measure of similarity between two sets of data by counting shared and distinct elements, are calculated to evaluate the strength of connections for all possible combinations of two words (Romesburg 2004). Given that causality lies in the network of words that are most strongly connected, a threshold value of Jaccard coefficient is set to only include words with strong co-occurrence in the network diagram.

This work utilizes an open-source linguistic software, KH Coder (Higuchi 2016), to construct the co-occurrence network diagram. Nodes are defined as target words with node size representing the word frequency and strength of edges are determined by the value of Jaccard coefficient between two nodes. KH Coder package employs a graph drawing method by force-direct placement to arrange the layout of networks (Fruchterman and Reingold 1991), and thus the graphical distance between words is irrelevant to its co-occurrence which is only indicated by edges between words. Certain words that are more closely associated with one another forms a community (or subgraph) coded by a certain color in the diagram. In the co-occurrence network diagram, several communities are observed, suggesting different types of events in the text data. Words within the same community are likely to carry a causal relation.

3.3.3 Workflow of NLP and text mining

Based on the methodology, the overall workflow of NLP and text mining of incident narratives is depicted as Figure 3.2 starting from preprocessing of narrative data to the end results of co-occurrence networks and K-means clusters. The preprocessing procedure follows the steps

explained in the previous sub-section. Attention should be paid to a customized filtering of general words relevant to the context yet not value-adding to new insights of contributing factors and causality, such as “pipeline”, “area”, and “station”. To determine the list of filtering words in the testing phase, a trial-and-error method is used by evaluating the results from K-means clustering and co-occurrence networks. When words appearing in the clusters or networks are not providing insights, they are listed in the filter, and the final list of filtering words are formed after a few iterations of the procedure. The effect of the filtering step is demonstrated in Figure 3.3 with word cloud of narrative data before and after the filtering. For example, words like “line”, “area” and “determined” are removed in the filtering step because they are not able to contribute to development of causation models. The words such as “release”, “operator” and “shut” that are more indicative of causality stand out in the word cloud after the filtering step. Admittedly, this manual step introduces subjectivity to the workflow, but it offers a fine-grained investigation on causality hidden in the text data. This filtering mechanism in nature is to apply domain expertise to helps resolve common difficulties in clustering analysis of unsupervised learning. The general words listed in the filter are provided in the supplemental material.

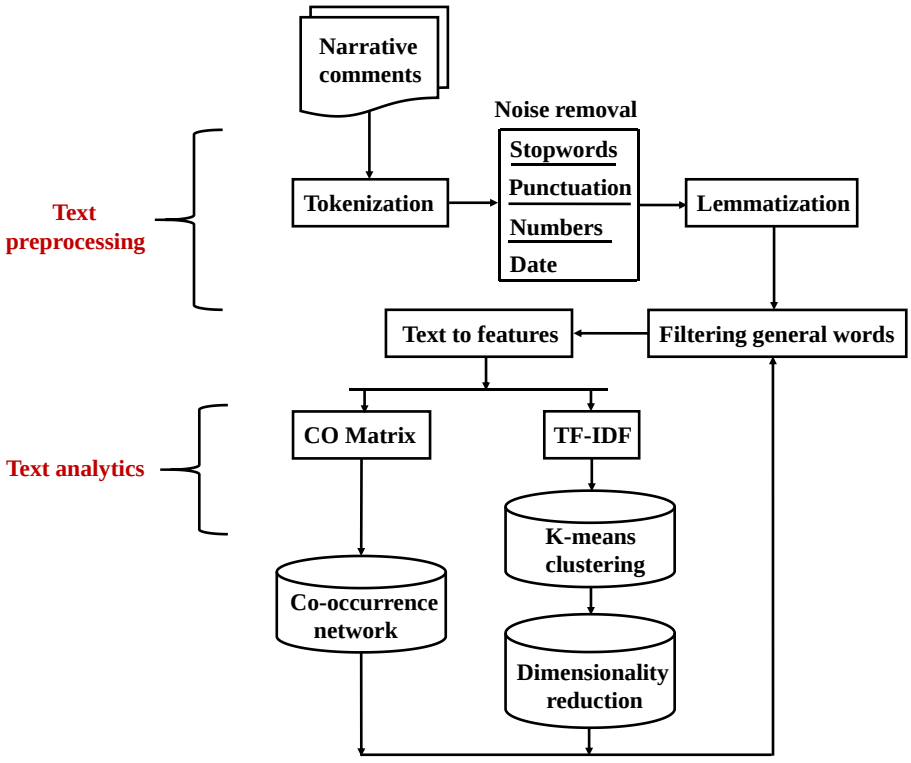

Figure 3.2 Overall workflow of NLP and text mining of incident narrative comments

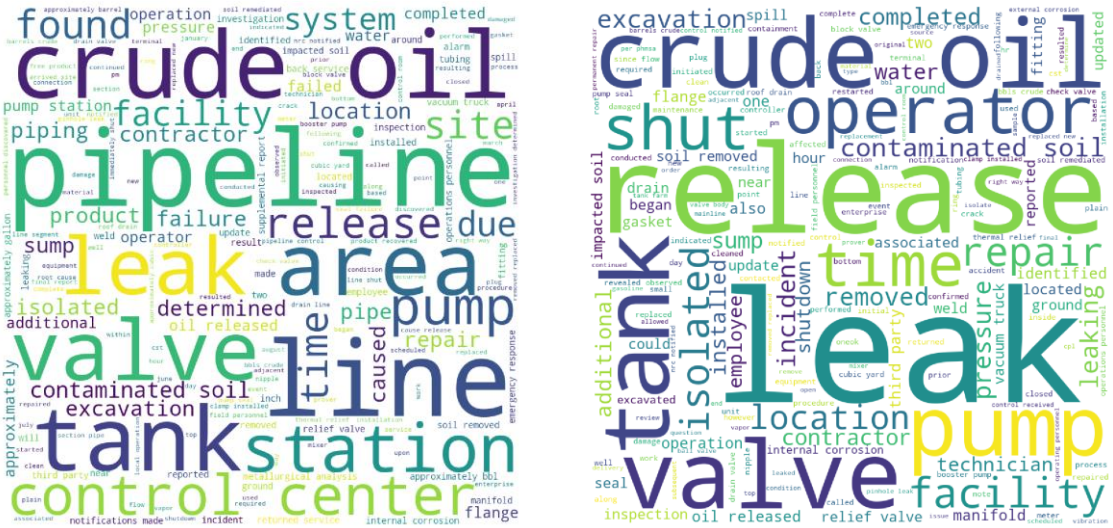

Figure 3.3 Word clouds of high frequency objects in the narratives before (left) and after (right) the filtering step

An extra step of dimensionality reduction applied after K-means clustering is to visualize and evaluate the clustering results by condensing the vectors into a two-dimensional representation. In this work, principal component analysis (PCA) as a linear technique and t-Distributed Stochastic Neighbor Embedding (t-SNE) as a non-linear technique are investigated. The comparison of results is discussed in next sub-sections.

3.4 Results Produced by NLP Using Incident Records

3.4.1 K-means clustering

K-means clustering is applied on TF-IDF matrix and produces clusters of documents with top words based on the accumulated TF-IDF scores. As determined by the aforementioned elbow method, the clustering results with K equal to five is explored. Meanwhile, evaluating K-means clustering algorithm is challenging due to the nature of unsupervised learning. Even though there is a considerable amount of research developing evaluation metrics (Halkidi et al. 2001, Amigó et al. 2009), the present work validates the clustering results by using techniques of dimensionality reduction to visualize the clusters of words. Both PCA and t-SNE are explored to transform high dimensional TF-IDF matrix into a two-dimensional representation and a comparison of the two methods is demonstrated in Figure 3.4. The result by t-SNE shows that the five clusters are well separated while a significant portion of overlapping is observed in the result by PCA. Apparently, t-SNE exhibits a strong advantage on visualizing nonlinear high-dimensional data over the commonly used PCA (Van der Maaten and Hinton 2008) and is applied in the present work. To

be noted, the clustering visualization produced by t-SNE may vary in terms of the shape and relative location of clusters when utilizing random initialization scheme due to its stochastic nature, but the five clusters are observed to be consistently well separated to demonstrate the validity of clustering results.

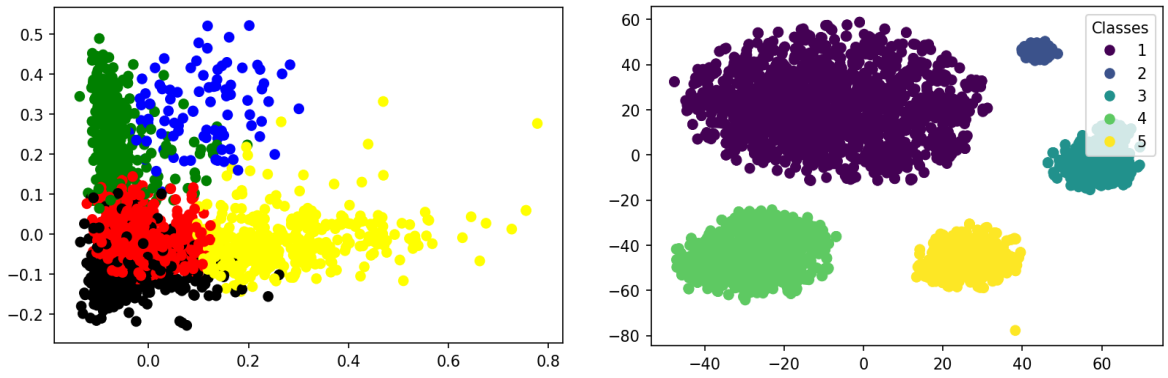

Figure 3.4 Two-dimensional visualization of clustering results with K = 5 by PCA (left) and t-SNE (right)

To visualize the clustering results, word clouds are constructed by the top 50 words in each of the five clusters shown in Figure 3.5 with the size of words based on the accumulated TF-IDF scores. There are certain levels of insights on contributing factors and causation indicated by this visualization. Considering Cluster 3 as an example, the words “leak” and “release” refer to the failure scenario, “contaminated” and “soil” are related to the consequence of the leak or release, “crude” and “oil” suggest the material involved, “internal” and “corrosion” can indicate the cause of failure which matches one of the seven pre-defined cause labels, the words “flange”, “valve” and “gasket” refer to the equipment where abnormality exists which refers to another cause label (equipment failure) defined by PHMSA, the words “notified”, “excavated” and “replaced” indicate the emergency responses, and “pressure” can be the contributing factor to internal corrosion or equipment failure. Thus, a cause-and-effect storyline of the pipeline incident can be constructed: the abnormal “pressure” leads to “internal corrosion” of pipeline “valves”, the “crude oil” is then leaked to the ground and “soil” is “contaminated”, and the emergency response team is “notified” to “excavate” the soil for remediation and the corroded valve is “replaced”. This induced causation can be validated by a sample incident caused by corrosion with reference # 20100005 in Appendix A as the aforementioned storyline captures the main events in the actual incident narrative. A detailed validation can be found in Section 3.4.3. Other clusters can be developed into similar storylines following the same approach. In Cluster 1, there are named entities like “suction”,

"vibration", "pump" and "booster" present together with "mechanical", suggesting mechanical failure of equipment as a potential cause, "crude oil" indicates the material involved, and "release" can be the consequence of pipeline incidents. Cluster 2 has words like "discharge", "overfill" and "overflow" refer to specific scenarios of incorrect operation and the presence of "thermal" can be induced as a cause factor of the aforementioned abnormal operating conditions. The top words in Cluster 4 are "leak", "release" and "valve" indicating the consequence, and "maintenance" and "drain" could be causes. Entities related to equipment appear in the top words of Cluster 5 such as "mixer", "tank", and "seal" which suggest where "leakage" and "release" occur, and the causal factors of "internal" and "external" "corrosion" can be identified. Thus, the K-means clustering results can provide insights on contributory factors of incident and its disadvantages are clear as well: a certain amount of manual interpretation and domain expertise is required to derive insights on the casual dependency from the clusters and it is still laborious to identify the words with casual relations from a number of associated words.

Figure 3.5 Word clouds of the total 3567 narratives developed by the top 50 words in each of the five clusters ranked by accumulated TF-IDF scores

3.4.2 Co-occurrence network analysis

Co-occurrence network has been employed by Syeda et al. (2017) to showcase the capability of NLP and text mining in safety research, but its application and analysis are hardly illustrated in detail. As K-means clustering can only produce clusters of associated words and fails to reveal word-word relations within the cluster, co-occurrence network is applied to overcome this limitation. The results of the total 3587 incident narratives are shown in Figure 3.6 with the threshold value of Jaccard coefficient as 0.18 (only including strong co-occurrence). The Jaccard coefficient is determined on a trial-and-error basis to include enough information in the diagram while still maintaining clarity of the diagram. The edge indicates whether co-occurrence exists between two nodes (or words). A group of words that possess strong co-occurrence with one another forms a community (or subgraph) in a specific color. Dashed edges suggest that words are of co-occurrence but in different communities. Unlike clusters generated by K-means clustering, the structure composed by nodes and edges in each community naturally forms a hierarchy of causation. Each community (or subgraph) with sufficient number of nodes typically leads to a storyline which describes a typical incident scenario. When interpreting the network results, subgraphs with too few nodes should be neglected and two connected subgraphs can also be merged into one storyline by appropriate interpretation.

The network structure in subgraph no.1 (in green) can construct an incident scenario: when an “operator” “closes” a “valve”, abnormal “pressure” is observed, and later “release” happens; three measures are taken (correspondingly three branches of edge are present) - safety “personnel” come to “isolate” the incident site, measurements are taken to “contain” the releasing gas, and “notification” is sent with an “estimate” of releasing “volume” and a report on “response” of this “emergency”. It should be noted that the subgraph no.1 is connected to subgraph no.2 (in yellow) via the node “personnel”, and the storyline can be expanded as the safety “personnel” come to not only “isolate”, but also “control” the “release”, “shut” down the “pump”, and “notify” relevant agencies. The development of this storyline requires less amount of manual work and is more straightforward following the linkage between nodes. Another subgroup no.3 (in Purple) probably represents that due to a “release”, “soil” is “contaminated” and is later “excavated” or “removed” for “recovery”. Or it can be as simple as that “internal” “corrosion” is identified as the major cause and the corroded part is “sent” to conduct “metallurgical” analysis as shown by subgroup no.6 (in Orange). It is also true for all the other communities in the diagram. Obviously, co-occurrence

exhibits strong performance in extracting contributing factors and revealing causality of incidents than K-means clustering. However, the co-occurrence network diagram of more than 3500 incident narratives can inevitably omit important contributing factors. For example, although "internal corrosion" is present in the subgraph no.6 of Figure 3.6, no further casual relations of corrosion are revealed.

To overcome this limitation, incident narratives under specific cause labels predefined by PHMSA are selected to construct co-occurrence network diagrams. Two illustrative cause labels are chosen: "corrosion" composed of 722 narratives with results shown in Figure 3.7 and "natural force damage" of 161 narratives with results shown in Figure 3.8. Due to the high word frequency, "corrosion" in Figure 3.7 is present in subgraph no.1 as a major node with more connected edges compared to its presence in Figure 3.6. A detailed cause-and-effect storyline of corrosion is unfolded following its connected nodes: "internal corrosion" occurs in the "tank" causing "release" of "crude oil", and "impact" of the incident is that "soil" is "contaminated" and then gets "excavated" to "recover"; meanwhile, as emergency response to the "corrosion", relevant "personnel" is "notified" and "dispatched" to the "field" to "control" and "shut down" the pipeline. The proposed storyline is similar to the one developed from Cluster 3 in Section 3.4.1 but requires considerably less amount of manual interpretation and domain expertise from practitioners. Similarly, fine-grained contributing factors are identified when using the narratives under the cause label of "natural force damage". Natural factors such as "hurricane", "flood", "heavy rain" and "lightning strike" are observed in Figure 3.8, which are omitted in Figure 3.4. The network diagram in Figure 3.8 can also be converted to a storyline following the connected nodes. Thus, the capabilities of automated content analysis of co-occurrence network are demonstrated with apparent advantage over unsupervised learning techniques such as K-means clustering.

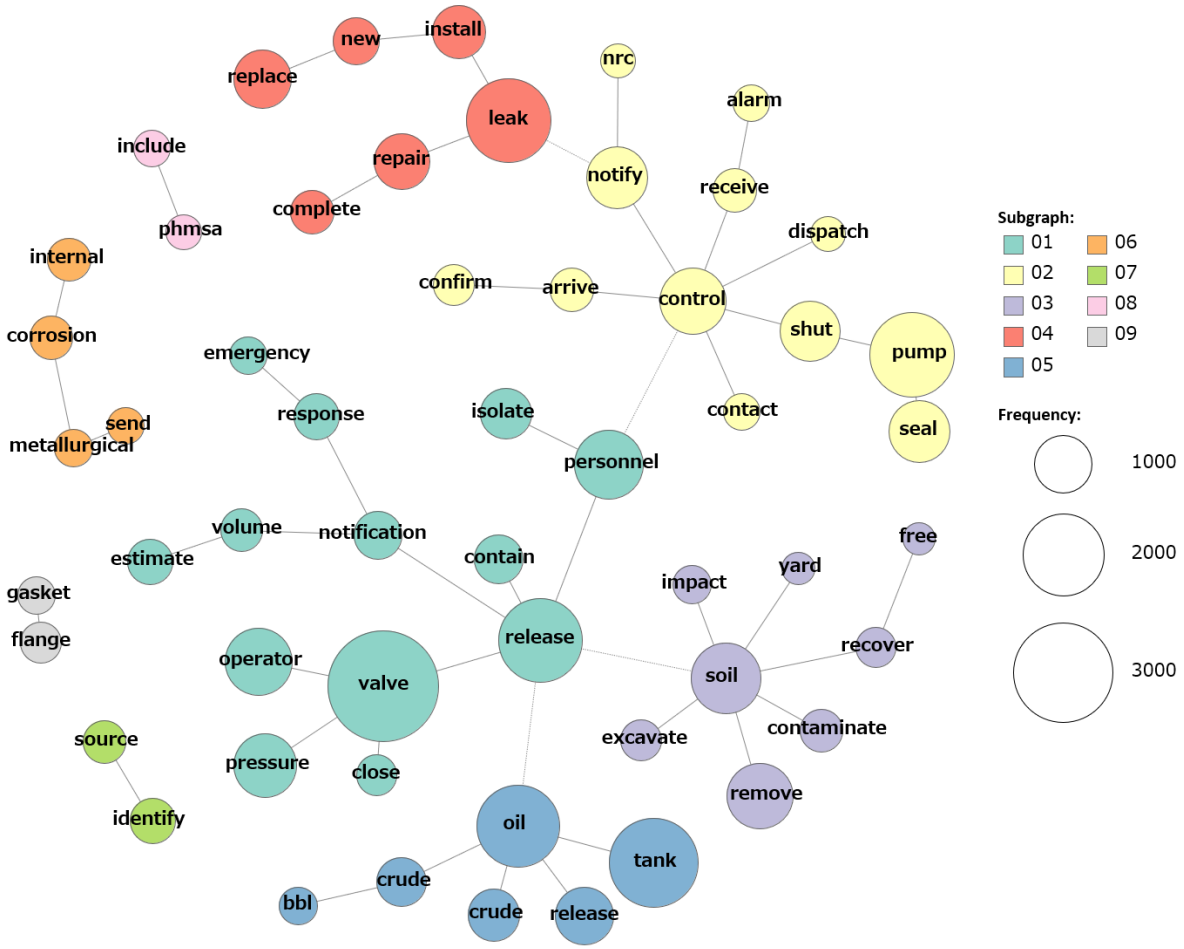

Figure 3.6 Co-occurrence network diagram of a total of 3587 incident narratives from PHMSA HL database

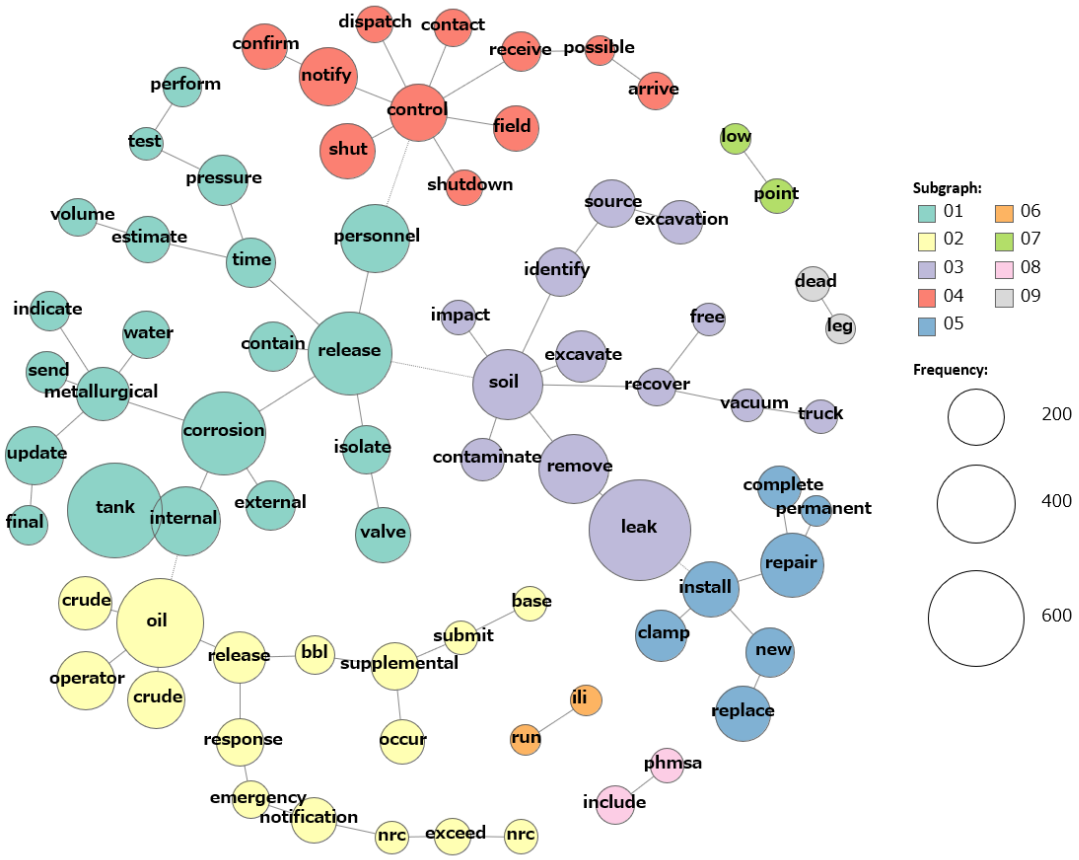

Figure 3.7 Co-occurrence network diagram of 722 incident narratives under the cause of "corrosion"

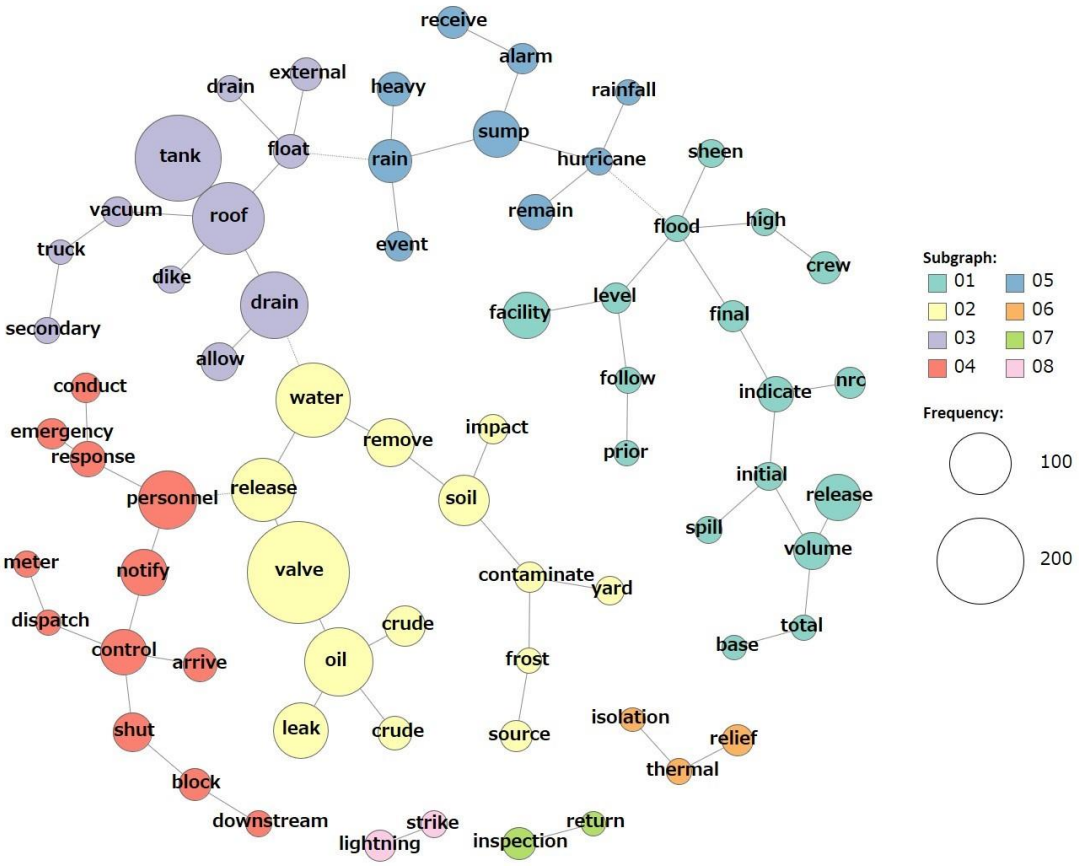

Figure 3.8 Co-occurrence network diagram of 161 incident narratives under the cause of "natural force damage"

3.4.3 Validation

While the co-occurrence network analysis of narrowed-down narratives under specific cause labels demonstrates strength of capturing latent dependency and causality of pipeline incidents, the network structure still comes from a large number of incident records that may have occurred from different scenarios, so the stories stated above may not be representative of all those incidents together (since each incident may have propagated in different ways). To validate the NLP-based analysis above, sample narratives are selected from the 3567 incident narratives as shown in Table 3.1. The authors perused the sample narratives and manually highlighted the key words (shown in bold) of incident descriptions. Several key objects in the samples such as "internal corrosion", "release", "crude oil", "soil", "excavated", and "pressure" are captured by the co-occurrence networks in Figure 3.6 and 3.7, and proposed storyline provides a straightforward way to describe the incidents. Key details in the samples with cause of natural force damage such as "lightning

strike", "drain" are omitted in Figure 3.6, but are captured in Figure 3.8 with a narrowed-down narrative dataset. However, the limitations of co-occurrence network analysis are also observed by comparing to the sample narratives: there are still some key information in the samples that are absent in the network diagrams, such as "power outage", "ammonia smell" and "evacuations"; (2) certain amount of domain expertise from practitioners are still required and a typical example is the interpretation of "contaminated soil" which could be mistakenly identified as a contributing factor of corrosion if the practitioners are not familiar with pipeline incidents; (3) the proposed storyline could be misleading as the network diagrams may miss key information, and as an example, "no other defect" and "defect" have opposite meaning, but with only "defect" present in the network, the constructed storyline can deviate from the fact.

Table 3.1. Sample incident narratives selected from PHMSA HL database (2010-2019)

Report Number	Cause	Narratives
20100005	**Corrosion** failure	**Internal corrosion** on 10-inch pipeline resulted in **release** of **crude oil**. **Spill** impacted an area measuring 20' x 30' x 7'. Did not occur on road, water, or ditch. Impacted **soils** were **excavated** and **remediated** on-site. Pipeline was inactive and upon **cleanup** activities, this segment of the pipe was **removed** entirely.
20100014	Corrosion failure	Lion oil called Sunoco control room to initiate a **delivery** to mid-valley pipeline at longview station. Sunoco control center advised Lion that when mvpl personnel completed the line up for delivery, Sunoco cc would advise Lion to start delivery. Lion appears to **have started actions** for the delivery **prior to being notified** by Sunoco cc that line up was complete. This resulted in a **higher than normal line pressure** at the Longview station manifold yet the pressure was within the design limits. This **higher pressure** appears to have caused the failure at the point where **internal corrosion** had occurred. Re-submitted on 3/19/2013 to include part e5f per phmsa request.
20100037	Material failure of pipe or weld	A report of **ammonia smell** was phoned into the city of Pawnee police department. Magellan operations control center was notified by the city of Pawnee police department at 6:22 pm Tuesday, January 12, 2010. The Skedee fire department and Magellan employees responded.

Report Number	Cause	Narratives
		Evacuations included 6 families. The **release** was located near/at mile post 58. The release was on the buried pipe. The section of the pipe where the release occurred was **removed** and **repaired** and the section was sent in for **metallurgical analysis**. Due to restrictions of the online form, question 14 "shutdown time" was omitted, but later added. The line was **shut down** at 18:25 upon notification for investigation. Release was verified at 19:08. This report was mailed 2/12/10 as the online reporting was not active.
20100081	Material failure of pipe or weld	On April 17, 2010 at approximately 11:30 am local time, while performing investigative follow-up work after a **brush fire** crossed lines 1 and 2, Enbridge environmental representative discovered and reported what appeared to be a small amount of **oil** on the right of way at mp 997.79. **Lines** 1 and 2 were **shut down** as a **precaution** and Enbridge pipeline maintenance personnel were **dispatched** to investigate. After hand-excavating the impacted area, a small **crack** was discovered and confirmed on the longitudinal seam of line 2 at approximately 7:00 pm local time. The **defect** was located in a marshy area and the site access and investigation progress was hampered by poor site access and ground conditions. External notifications to the national response center and Minnesota state duty officer were made upon confirmation of the **leak**. Notifications were also made directly to the Minnesota office of pipeline safety, phmsa, Minnesota pollution control agency and the Minnesota interagency fire center. Once the pipe was **excavated** the entire long seam of the joint of pipe in question was field assessed by nde (both ut and magnetic particle inspection). **No other defects** were identified. An integrity assessment was completed by Enbridge's pipeline integrity group and the pipeline was **repaired** using a plidco split sleeve. The return to service plan was reviewed with phmsa and mnops and line 2 was restarted on April 18 at approximately 7:10 pm. The integrity assessment of this line segment is ongoing and the section of pipe in question will be cut out for further **metallurgical analysis** when conditions allow. Enbridge environment group is managing site **cleanup**

Report Number	Cause	Narratives
		and **restoration** in conjunction with a number of agencies (both local and state). A metallurgical analysis was conducted and the **leak** was found to be the result of a **hook crack** that was formed at the time of pipe manufacture which subsequently extended by **fatigue** through the remaining thickness of the pipe. there was **no evidence** that was found to suggest that either **post manufacturing mechanical damage** or **corrosion** had contributed to the flaw responsible for the **leak**. The amount of **contaminated soil removed** from the leak site was 30 cubic yards.
20100039	Equipment failure	The **spill** was a result of a **crack** in the **flange** of an existing cast iron **valve**. The cast iron valve was **bolted** to a cast steel flange. The old cast iron valve was **replaced** with a cast steel valve.
20100166	Equipment failure	The location of the **release** is a **pump** station owned and operated by TransCanada keystone pipeline, lp (TransCanada). The **pump** station is in a rural area located at approximately three miles south of Roswell, south Dakota. The **release** of **petroleum** was entirely contained on TransCanada property. The **release** occurred from a **loose fitting** on an above ground damper system associated with an injection pump. **Oil** was **released** from the **loose fitting** for an approximate 3 second period until the system was manually **shut down**. An estimated 100 gallons of oil sprayed over an area of approximately 60 feet by 110 feet within the pump station location. TransCanada personnel were onsite at time of the oil release, the injector pump was immediately **shut down** and **containment** and **recovery** activities initiated. A maintenance team mobilized to the site upon notification of the release on June 23, 2010 at 12 noon CDT. The pulsation dampener on the injection pump was **removed** and visually inspected. The inspection revealed the **threaded nipple** was **not installed properly** and was **not the correct length**. The threaded nipple on the pulsation dampener was **replaced**.
20100189	Natural force damage	A **lightning strike** caused a **power outage**, causing the mov to close, which resulted in the **relief valve opening** and **overfilling** the **sump**. Approximately seven gallons of gasoline were **released** and seven

Report Number	Cause	Narratives
		gallons **recovered**. Impacted **soils** were place in drums and will be hauled off site to an approved facility.
20100026	Natural force damage	Tank 824 water drain was **leaking crude**. Due to **extreme temperatures** in the area, it is believed the roof **water drain piping froze** and compromised the drain piping integrity. When t824 was filled with crude, the **frozen** components **thawed** allowing product to exit the **tank** via the **drain piping**. As of 4-7-2010, t824 is still in service. Once **removed** from service, the exact cause can be determined. As of 5-25-2010, t824 had been taken out of service and **cleaned**. It was determined that the roof drain hose integrity had been compromised in 2 locations due to **ice expansion**. Updated per blaine keener e-mail due to changes in PHMSA reporting form.

3.5 Results produced by NLP using Incident Investigation Reports

3.5.1 Cooccurrence Network Analysis

Recognizing the significant influence of the quality of data sources (*i.e.*, incident reports) and necessity of appropriate supervision in the workflow on the output of analysis, the cooccurrence network is observed with significant improvement by two modifications:

- feeding a subset of collected incident reports preferably under the same major cause and even sub-cause (instead of feeding a large number of reports with miscellaneous causes).
- specifying keywords (*i.e.*, the center of subgraphs in the cooccurrence network diagram) chosen using empirical knowledge (instead of top words of frequency) as the role of supervision that helps build the cooccurrence network structure toward a more explicit causal representation.

Cooccurrence network diagram of 9 incident reports under the cause of corrosion shown in Figure 3.9 indicates noticeable improvement on the readiness of the network to be used for automatic extraction of contributory factors of pipeline failure. The top five words of frequency is by default used as keywords, resulting in 5 subgraphs in the diagram with each key word positioned in the center. The “investigation” subgraph linked with other two subgraphs indicates that identification of the incident “cause” as “corrosion” and “leak” at certain “location”. The sub-causes (or

contributory factors) are clustered around “corrosion” – including “mic” (*i.e.*, microbiologically influenced corrosion), “inspection”, “internal” and “external”. The subgraph of “leak” is connected to “tank” and “pipeline” subgraphs which are two major equipment where leak mainly occurs.

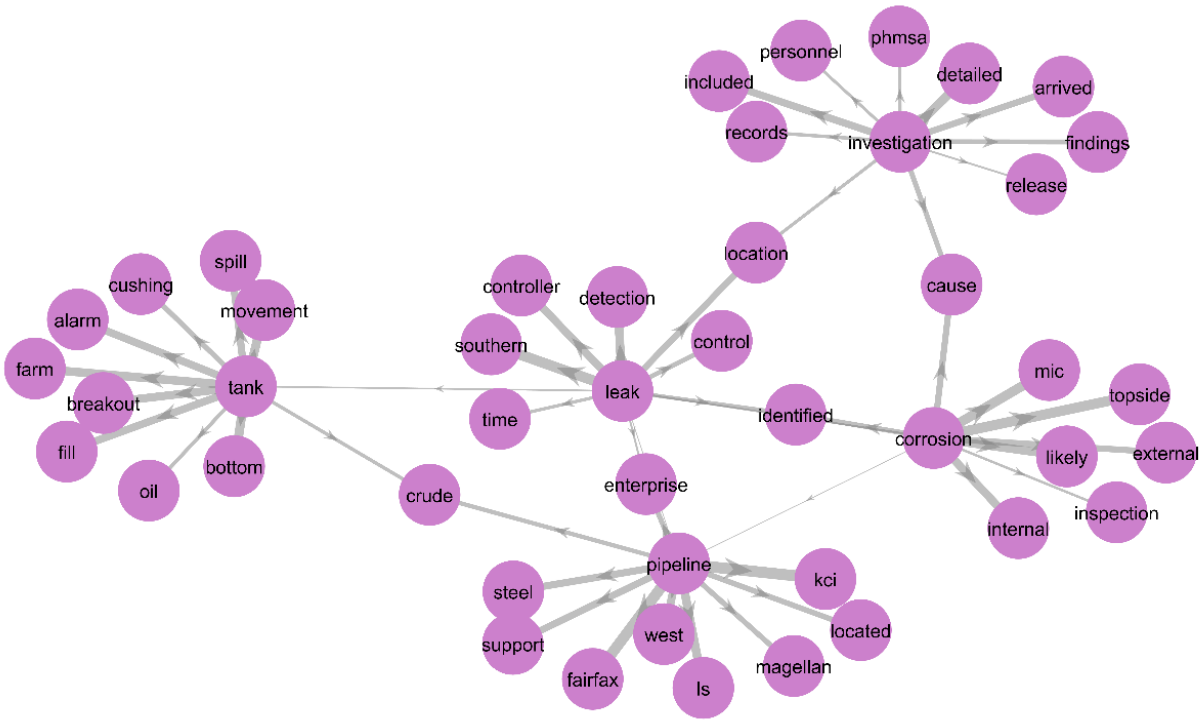

Figure 3.9 Cooccurrence network diagram of 9 corrosion incident reports (a total of more than 15000 words) with five keywords by default chosen as most frequent words

However, with keywords chosen rarely by word frequency, certain subgraph may not contain valuable information such as the "investigation" cluster that is not directly related to the cause-effect relation of the pipeline failure. Thus, cooccurrence network diagram of 9 corrosion reports with keywords specifically selected from the word frequency list: "corrosion", "leak", "tank", "pipeline", and "internal" is shown in Figure 3.10. The subgraph of "internal" is connected to the "tank" subgraph via "bottom", suggesting the location of internal corrosion in the tank. All the subgraphs are linked with cause-effect relation, contributed to the supervising step of manual selection of keywords in the workflow.

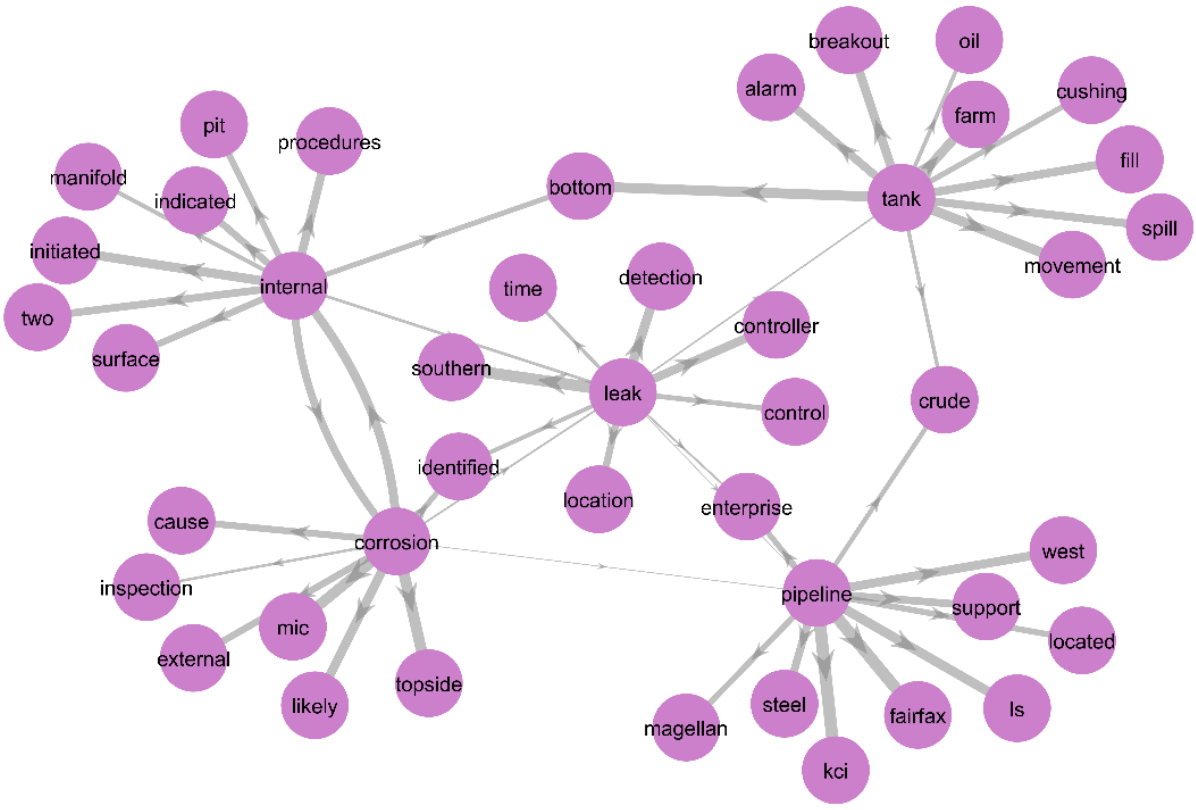

Figure 3.10 Cooccurrence network diagram of 9 corrosion incident reports with five keywords selected with empirical knowledge

The aforementioned two modifications are further evaluated by feeding a more targeted data source. According to PHMSA, corrosion incidents can be classified as (1) corrosion in the tank, and (2) corrosion in the pipeline. The 9 incident reports under the major cause of corrosion are then split into one set of 4 corrosion-in-tank reports and the other set of 5 corrosion-in-pipeline reports. Cooccurrence network diagram of the "tank" reports with default keywords is shown in Figure 3.11 where "crude" and "oil" subgraphs are overlapped, and the overall network structure is not well-connected to formulate the causality of incident. An improvement is observed in Figure 3.12 using the specified keywords (*i.e.*, "corrosion", "leak", "tank", "pipeline", and "internal"), which demonstrates the role of supervision on reshaping the cooccurrence network structure. The analysis of 5 "pipeline" reports is presented in Figure 5 and 6. While the network structure seems well-linked in Figure 3.13 without supervision, subgraphs of "enterprise" and "magellan" are not informative to derive causal factors. Figure 3.14 exhibits a greater level of information density by specifying keywords directly related to the pipeline such as "valve" and "pressure". Thus, semi-

supervised workflow with supervision of specifying keywords (or “center” words) and data source selection is found to significantly improve the NLP analysis of incident causality.

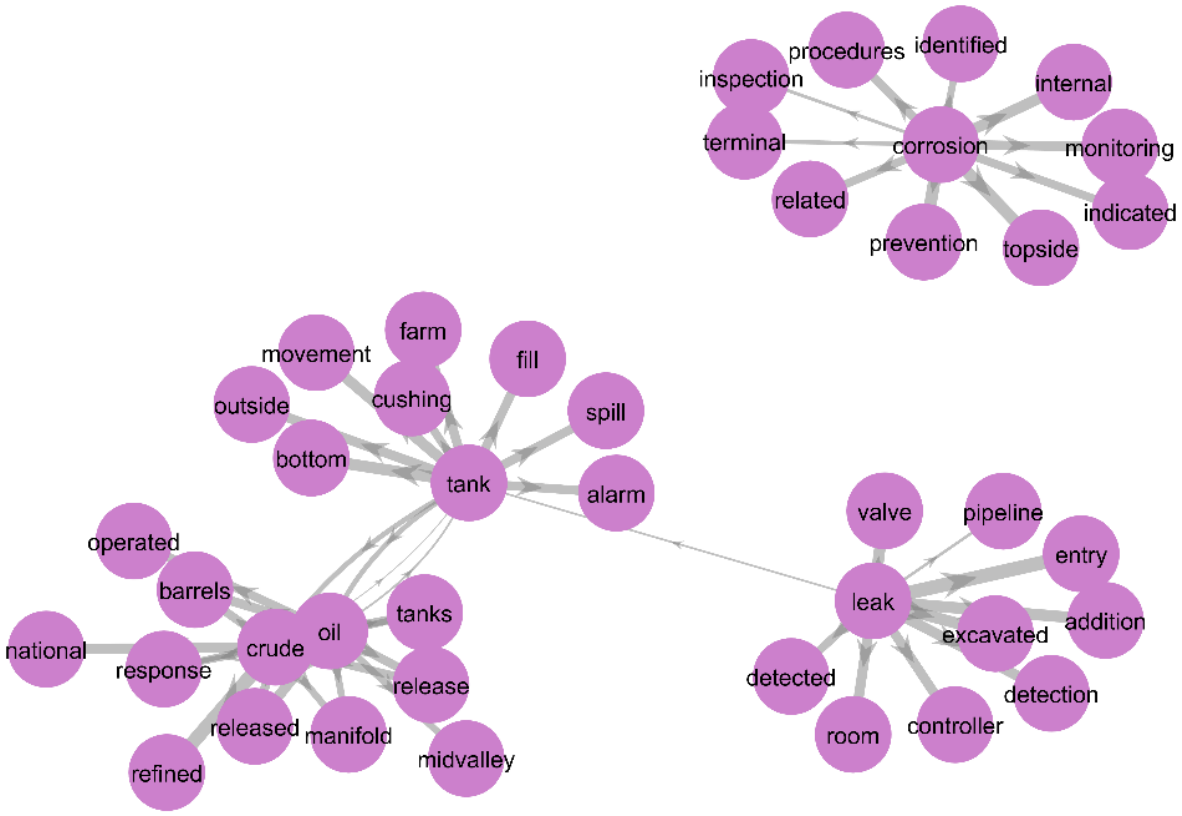

Figure 3.11 Cooccurrence network diagram of 4 incident reports of corrosion in the tank (a total of more than 6000 words) with five keywords by default chosen as most frequent words

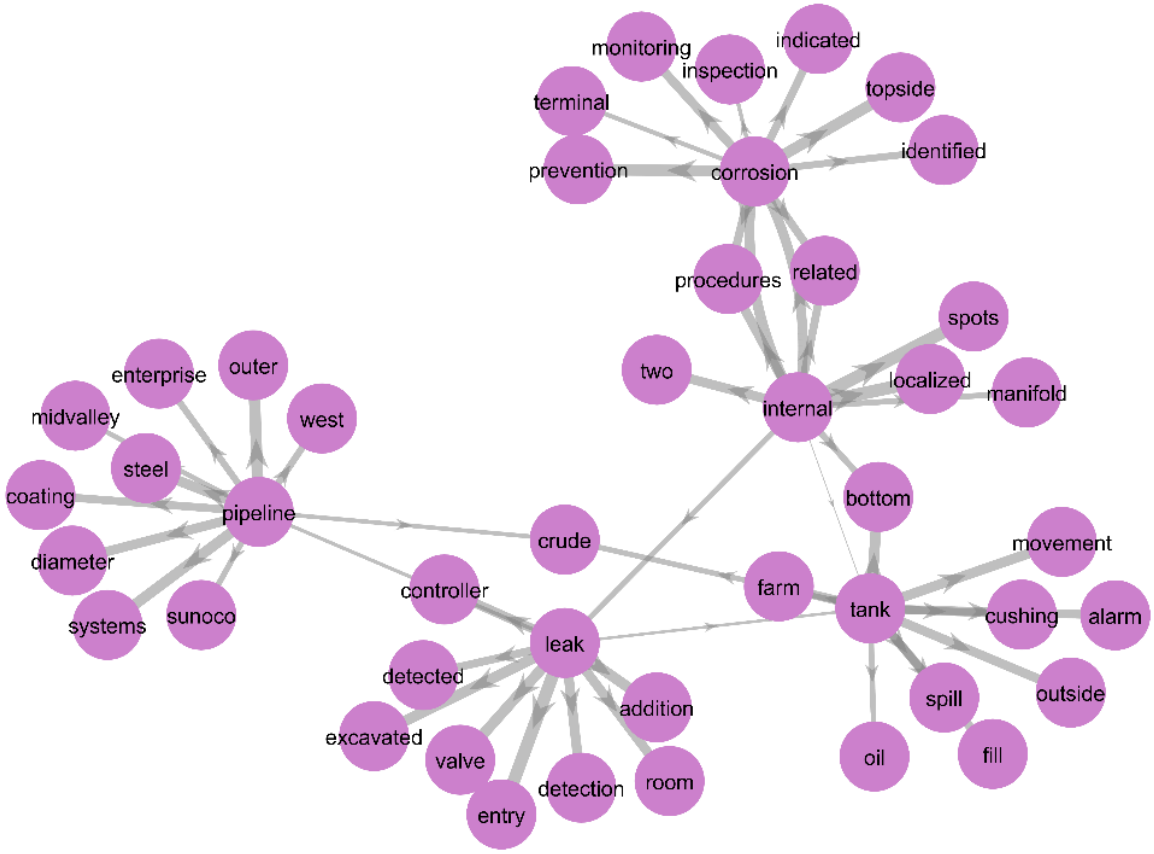

Figure 3.12 Cooccurrence network diagram of 4 incident reports of corrosion in the tank with five keywords selected with empirical knowledge

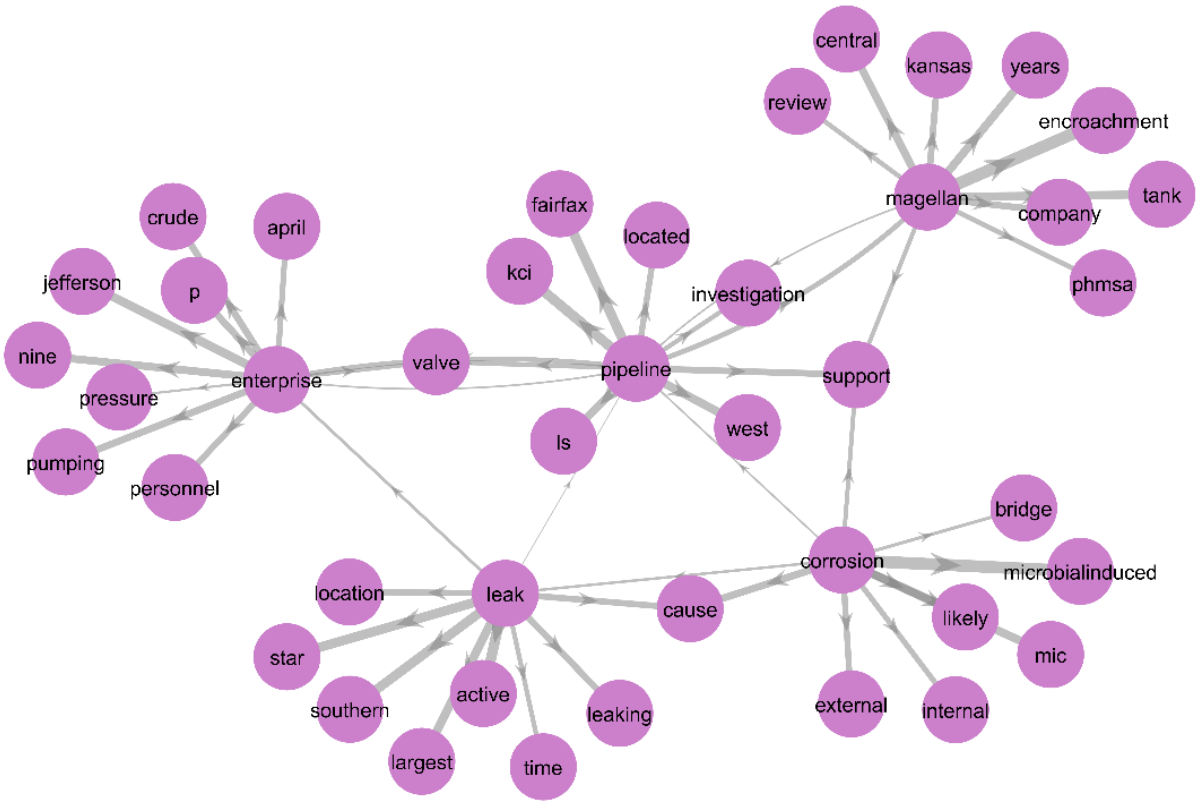

Figure 3.13 Cooccurrence network diagram of 5 incident reports of corrosion in the pipeline (a total of more than 9000 words) with five keywords by default chosen as most frequent words

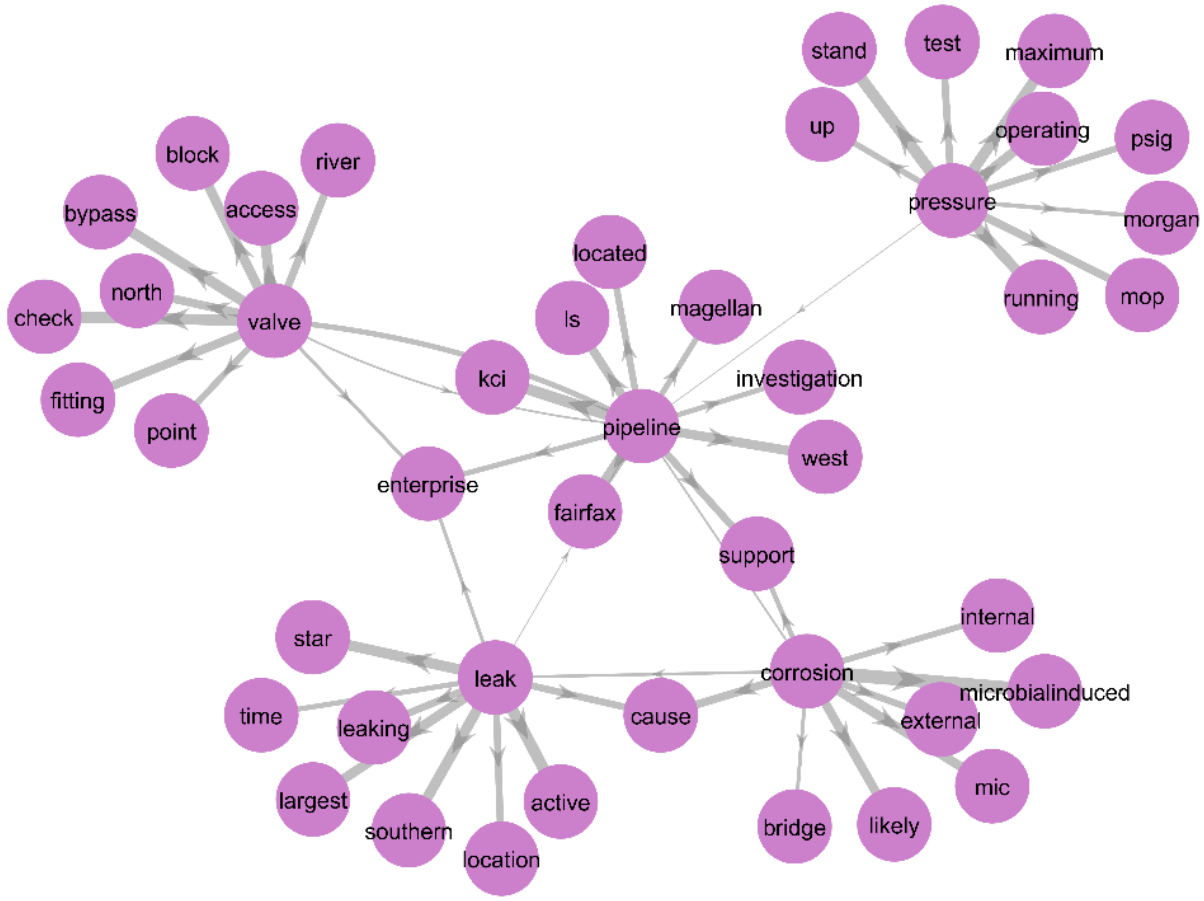

Figure 3.14 Cooccurrence network diagram of 5 incident reports of corrosion in the pipeline (a total of more than 9000 words) with five keywords selected with empirical knowledge

3.5.2 Topic Modeling Analysis

The topic modeling with LDA is applied to 9 incident reports of corrosion, a subset of 4 reports of corrosion in the tank and the other subset of 5 reports of corrosion in the pipeline. All the results show (Figure 3.15, 3.16, 3.17) that the topic modeling is able to identify key information of failure, but the assumption of neglecting word sequence impedes generating more insights on cause-effect relation.

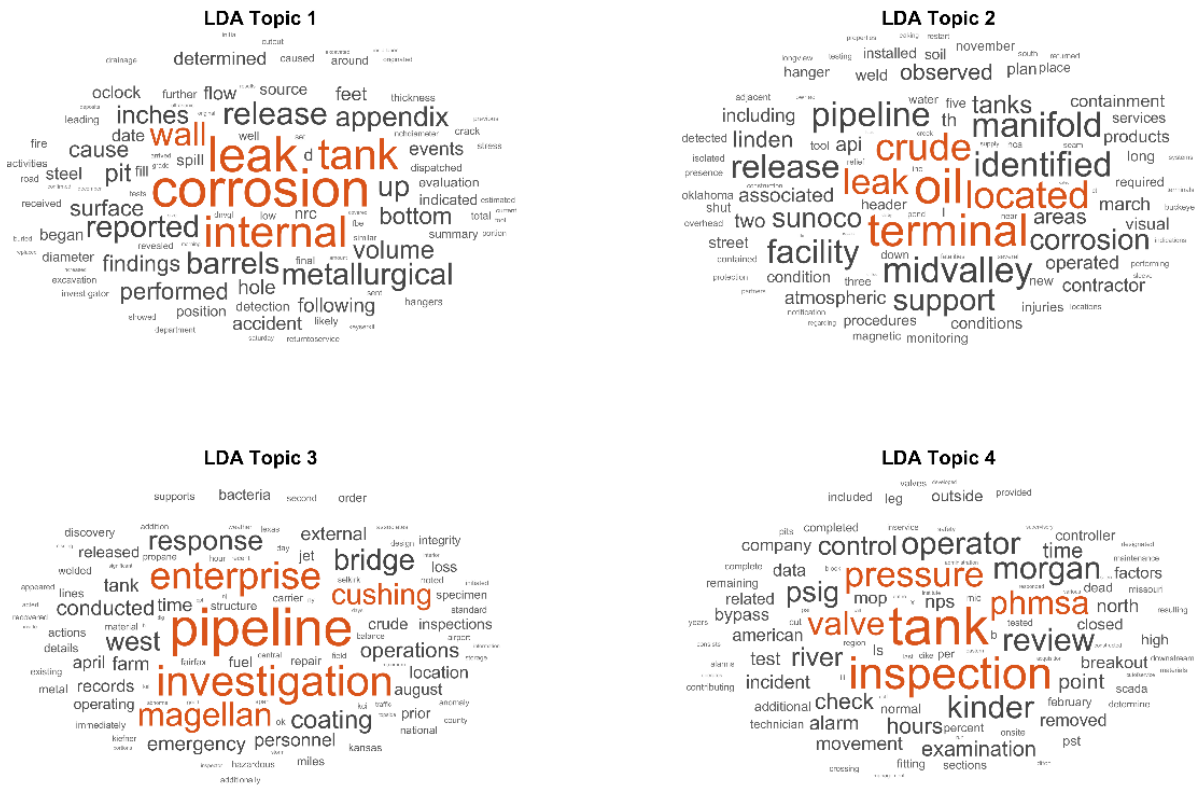

Figure 3.15 Topic modeling with LDA applied to 9 corrosion reports

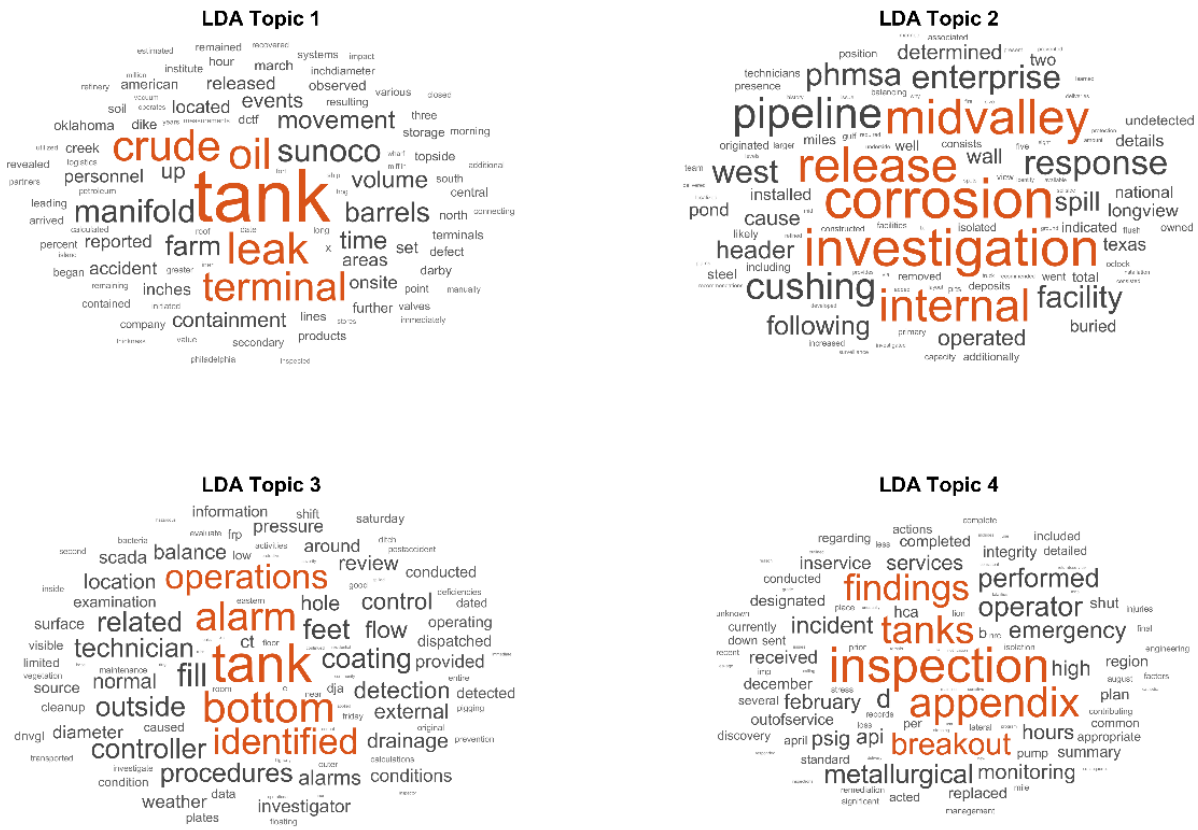

Figure 3.16 Topic modeling with LDA applied to 4 reports corrosion in the tank

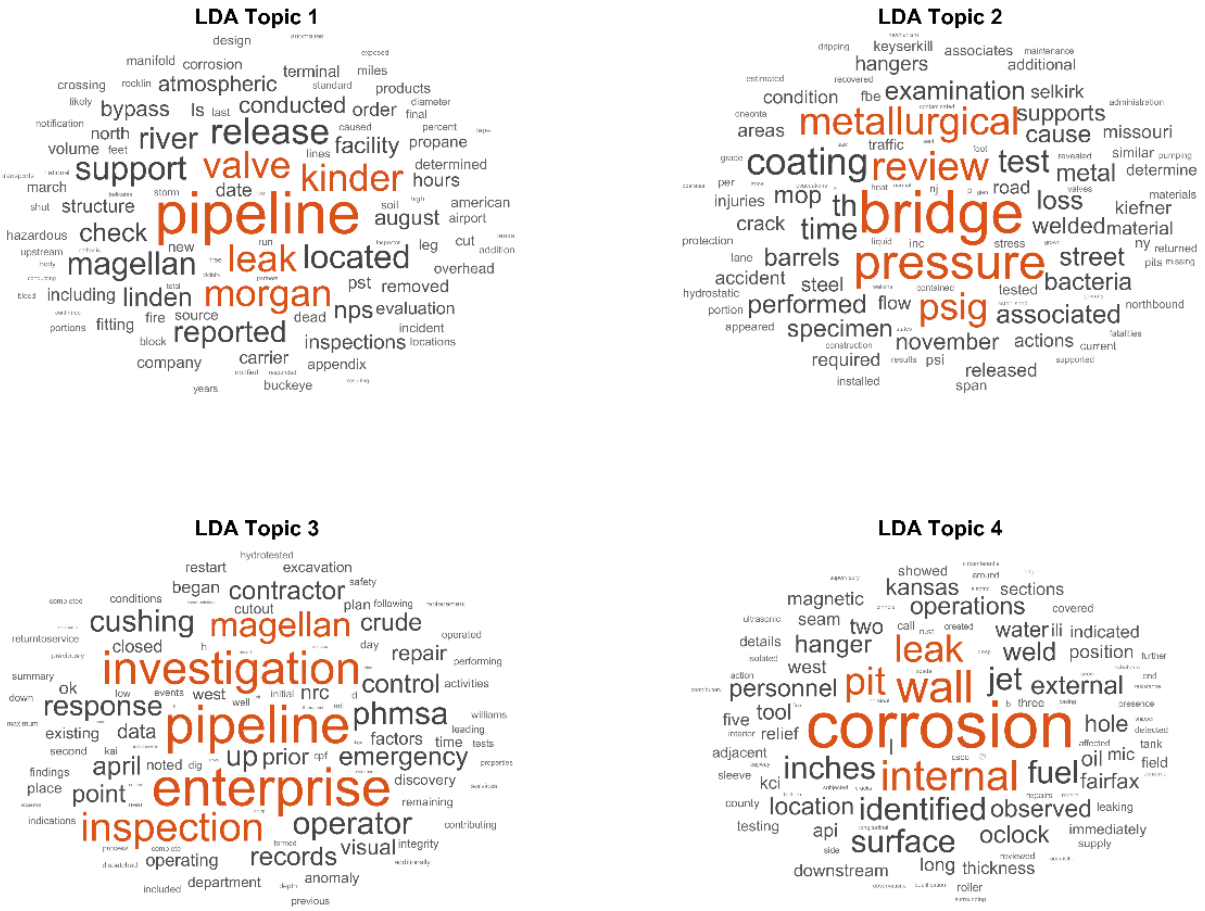

Figure 3.17 Topic modeling with LDA applied to 5 reports corrosion in the pipeline

3.6 Taxonomy for Causal Analysis

PHMSA has established a comprehensive mapped cause system that defines the primary or direct causes of the incident. As described in the section 2 that it has seven primary causes: corrosion, material/weld/equip failure, excavation damage, incorrect operation, natural force damage, other outside force damage, and all other causes. Each of these categories has some sub-categories which are also well-understood. Beyond these two tiers PHMSA provides opportunities to the operator to provide more information through various fields in the incident reporting system such as pipeline specification (pipe diameter, pipe thickness, material of construction, year of installation), operational specification (pressure, commodity transported, depth of cover), inspection and maintenance detail (date of last inspection, inspection methods, prior incident), incident detail (date, release amount, asset damage, environmental damage), and incident narratives. This information is very valuable and important. In next section, they have been utilized to develop an ANN model to predict the pipeline failures. Some of these provided datasets are termed as

background factors that have no apparent contribution to the failures. But they do have specific relations to the failure data as observed from the literature review and the data presented in previous section. For instance, smaller diameter pipeline has no specific reason to fail unless they are subjected to a specific operational condition. PHMSA incident records did a brilliant job in collecting such background factors over the time. However, they need no to be confused with the underlying causes that can lead to an incident. For instance, inadequate maintenance can alone cause an incident, but pipe diameter cannot. There is a need to include such underlying causes in the incident records for all recorded incidents. The mechanism to do so will need further understanding and careful examination. A set such underlying causes has being used by NEB for a while. They are well described in the previous section and also given in the table below. There are different sets of information that can be helpful as well to better understand the failure conditions.

Mapped Causes and Sub-causes
Corrosion: Internal corrosion, External corrosion (General corrosion, localized pitting, galvanic corrosion, atmospheric corrosion, stray current corrosion, microbiological corrosion, selective seam corrosion, others)
Material/Weld/Equip Failure: Construction, installation or fabrication-related, defective or loose tubing/fitting, environmental cracking-related, failure of equipment body, malfunction of control/relief equipment, manufacturing-related, non-threaded connection failure, other equipment failure, pump or pump-related equipment, threaded connection/coupling failure
Excavation Damage: Operator/ contractor excavation damage, previous damage due to excavation, third party excavation damage
Incorrect Operation: Damage by operator or operator's contractor, incorrect equipment, incorrect installation, incorrect valve position, other incorrect operation; overfill/ overflow of tank/ vessel/ sump, pipeline/ equipment over-pressured
Natural Force Damage: Earth movement, heavy rains/floods, high winds, lightning, other natural force damage, temperature
Other Outside Force Damage: Electrical arcing from other equipment/ facility, fire/ explosion as primary cause, fishing or maritime activity, intentional damage, maritime equipment or vessel adrift, Other outside force damage, Previous mechanical damage, Vehicle not engaged in excavation
All Other Causes: Miscellaneous, Unknown

Underlying Causes
Engineering and Planning: failures of assessment, inadequate planning or monitoring, inadequate specifications or design criteria, lack of evaluation of change, or implementation of controls
Maintenance*:* inadequate preventive maintenance or repairs, failure to maintain excessive wear and tear
Inadequate Procurement**:** failures in the purchasing, handling, transport, and storage of required materials
Tools and Equipmen*t*: improper use or inadequate tools and equipment
Standards and Procedures: inadequate development, communication, use, maintenance or monitoring of standards and procedures
Failure in Communication*:* loss of communication with automatic devices, equipment, or people
Inadequate Supervision: lack of oversight of a contractor or employee during construction or maintenance activities
Human Factors*:* individual conduct or capability, or physical and psychological factors,
Natural or Environmental Forces: external natural or environmental conditions
Background Factors

4.0 ACIDENT MODELING USING ANN

4.1 Objective

Once it is possible to convert the root cause analysis into suitable extractable data, the next task is to build an ANN model. A neural network needs to be developed by training. This refers to using known input-output pairs as examples to teach the model to determine the weights between the connecting neurons. Since a knowledge-based model that learns from past incidents is the overall objective of the ANN, inputs can be those deviations that had contributed to the output, which are the consequences of the incidents.

Records of inspection and maintenance data, laboratory testing and failure analysis can be used in conjunction with the past incident records. Past incidents may have occurred in other facilities, or locations under different management system and this makes data from past incident records generic. Laboratory findings or inspection records for example may contain information specific to a site and should not be overlooked just because a large number of past records are available. Thus, all information should be fed to the neural network to train it to better predict failure. The challenge lies in determining what information should be fed to the training model along with those identified in Task 1.

Once it is determined what information are to be used for training, the task lies in developing the actual model via training. Various literature exist that suggest different methods of determining the suitable size of ANN model and proposes learning algorithm to determine the connection weights that lead to a tolerable error (Lappas 2007, Nuchitprasittichai and Cremaschi 2013, Cortes et al. 2016) Joghataie *et al*., 1995]. The best suitable size and algorithm for model will have to be selected and training conducted based on findings.

To train the neural network so that it determines the non-linear relationships between the various causes and the consequences, past incident investigation records will be used as training examples. If information about pipeline failures can be gathered, 2/3 of the data can be utilized for training the network, while the remaining 1/3 can be used later for validation.

Once the model has been validated, it can now be used for prediction. Information about current conditions can be fed to predict the failure probability of pipelines. As new incidents occur, the information from them can be constantly used to update the ANN so that better prediction is achieved as time goes by.

4.2 Current Approach for using Artificial Neural Network (ANN)

Pipelines are one of the safest modes to transport bulk energy and have failure rates much lower than railroads or highway transportation (Carvalho et al. 2008). Yet pipeline failures do occur, sometimes with catastrophic consequences (Guo et al. 2016). An accurate risk analysis of pipeline incidents can result in an effective prediction of how various conditions contribute to an increased risk of such incidents and allow strategic measures to be developed for management of the overall risk of pipeline incidents.

The causes of pipeline incidents can be broadly classified into five categories: corrosion, equipment failure, natural force, operational error, and third party induced damage (Dey et al. 2004, El-Abbasy et al. 2014). Corrosion can be further categorized into internal corrosion which is influenced by the internal environment of a pipeline such as material being transported, and external corrosion affected by factors such as pipeline coating, cathodic protection measures and other factors. Incidents due to equipment failures consist of cracks and fractures that are unable to withstand the pipeline flow, and those due to natural force are caused by events such as floods, earthquakes, snowstorms, etc. Incidents due to operational error are those that are influenced by fluctuations in operating conditions (e.g., pressure), and third-party incidents represent damages caused by an operation not carried out by the pipeline operator itself (e.g., excavation done by contractors). Among all these causes, corrosion failures comprise about 25% of onshore hazardous liquid (Banimostafa et al. 2012), transmission pipeline incidents(Halim et al. 2020), and hence, they are ranked as one of the most frequent cause of HL pipeline incidents (Muhlbauer 2004, Davis et al. 2006). Effective risk management measures would require a detailed risk analysis of corrosion-induced pipeline incidents for informed decision-making purposes.

In the USA, data report of a pipeline incident is submitted to the Pipeline Hazardous Material Safety Administration (PHMSA) by pipeline operators within 30 days of an incident (Lam and Zhou 2016). The key data fields collected for the incident contain in-depth information about location, facility, operating conditions, number of injuries and/or fatalities, commodity released, causes of failure, etc. The total number of data fields collected by the PHMSA for an incident is 606. The high number of reported data fields in the PHMSA database indicates the detailed information gathered for each incident and presents an opportunity to leverage the detailed data for an accurate causal, consequence, probability, and risk analysis of pipeline incidents.

Established methods of risk analysis in the pipeline industry rely on using statistical trend analysis of past incidents to assess and predict the causes, consequences, and probability of pipeline incidents, which are combined to estimate the risk of pipeline incidents (Papadakis 1999, Lam and Zhou 2016, Bubbico 2018). Specifically, historical incident databases have been analyzed to derive the most frequent cause, the average rate of injury and fatality, and the average rate of incidents to quantify cause, consequence, and probability of pipeline incidents, respectively. However, a statistical trend analysis without a reasonable understanding of the interplay among key contributors of an incident cannot provide a clear understanding of the risk (i.e., cause, consequence, and probability) of a pipeline incident. Valuable information is lost in such superficial analysis which could otherwise guide decision-makers to identify the issues that drive pipelines towards higher incident rates and how these can be managed to reduce the risk of pipeline incidents.

For a deeper understanding of the risk of pipeline incidents, data-based models have been developed utilizing methods such as neural network, regression technique, and Bayesian methods(Breton et al. 2010, Senouci et al. 2014, Senouci et al. 2014). However, existing causal and consequence models utilizing these methods use only a small number of data fields for prediction of cause and consequence of pipeline incidents, in spite of the presence of hundreds of data fields, thereby losing valuable insights (Najafi and Kulandaivel 2005, Li et al. 2016, Mazumder et al. 2021). Additionally, in the field of data-based incident probability estimation, it is commonly assumed that failure rates are constant (does not change with time), and homogeneous Poisson processes are utilized to consider a constant average failure rate (Restrepo et al. 2009, Shan et al. 2018, Carpenter et al. 2019). However, changes in the system brought about by multiple contributing factors together alter the failure rate over time and an assumption of constant failure rate becomes invalid. More accurate incident probability estimation calls for models that can adjust the failure rate based on gathered data for better probability prediction of a pipeline incident. Furthermore, although integrated models for causal, consequence, and probability analysis of pipeline incidents are present in literature, they are only applicable for risk estimation utilizing the historical incident data. In other words, they are not equipped to utilize the current condition of pipeline conditions and predict the risk of a pipeline incident in the near future.

To handle the above-mentioned limitations in risk (i.e., cause, consequence, and probability) analysis of pipeline incidents, the objective of this article is to develop an integrated risk prediction model for pipeline incidents. Specifically, the proposed model presents a framework for the prediction of likely causes of corrosion-induced pipeline incidents, the subsequent consequences, and the incident probabilities. The causes and consequences of a corrosion-induced pipeline incident are learned using a machine learning approach whereby significant data fields from the rich pipeline incident database are utilized to develop artificial neural networks (ANNs). Next, to predict the probability of pipeline incident, a nonhomogeneous Poisson process model which considers varying failure rates is utilized through a Bayesian analysis. While utilizing the proposed framework for risk prediction, the information about the current condition of the pipeline is fed into the ANN models and Bayesian analysis to predict the cause, consequence, and probability of corrosion-induced pipeline incident, thereby giving a comprehensive look at the risk of future incidents.

This section is organized as follows. First, detailed information about the pipeline incident data utilized for the risk prediction and its preprocessing method is provided in Section 4.3. Next, the proposed methodology that consists of ANN models and Bayesian analysis is explained in Section 4.4, followed by a demonstration of the proposed risk analysis framework on corrosion induced pipeline incidents in Section 4.5.

4.3 Data Processing

In North America, the oil and gas pipeline incident database is managed by the PHMSA. Pipeline operators are required to report every event that involves an undesired release to the environment and meets any of the following criteria (Bolt et al. 2006) to the PHMSA:

1. The incident involves a death or personal injury resulting in hospitalization
2. Estimated property damage including the cost of commodity lost is greater than $50,000.

In this work, the data has been collected from the PHMSA database corresponding to the onshore HL transmission pipelines in the US between 2010 and 2019. The collected data has 3,592 pipeline incidents, and each pipeline incident has 606 data fields. One of the most frequent causes in onshore HL transmission pipeline incidents in the last 10 years has been corrosion with 721 incidents recorded over this time in the database. To develop a risk (i.e., cause, consequence, and probability) prediction model for corrosion-induced pipeline incidents in onshore HL transmission

pipelines, the data for corrosion-induced pipeline incidents is preprocessed for an effective risk prediction.

4.3.1 Data for cause and consequence prediction

For a cause and consequence prediction model for corrosion-induced pipeline incidents, 70 out of 606 data fields are initially selected based on reasoning about their significance to corrosion-induced pipeline failure. There are two types of data fields among the selected ones: (a) generic data fields relevant to a failure (e.g., time, location, and area of the incident), and (b) data fields specific to corrosion failure (e.g., presence of corrosion inhibitors and lining). The challenge lies in the sparsity of the information in certain data fields and some level of aggregation is thus required while ensuring that the granularity of the data is not lost. In other words, some of these data fields are populated for only a small number of incidents. Therefore, these data fields have been combined to increase the information density of data fields. For example, among generic attributes, age of pipe and age of tank have been combined to account for the age of the item involved in the incident. As an example, for corrosion-specific attributes, there are n data fields related to inspection types, with each field representing a different type of inspection (such as magnetic flux, ultrasonic and triaxial inspection). Each of these fields has 1 and 0 as its values, depending on whether it has been performed or not. These n data fields have been combined to result in a new data field with its values as 1 to n depending on the recent inspection type performed. In this manner, the number of selected data fields has been reduced from 70 to 26.

Numerical operations have then been performed on the selected data fields to produce more informative data fields. For example, the difference between the data fields, the year of manufacture of the item and the year of the incident, is utilized as the age of the item involved in the incident. Additionally, since most of the data fields (e.g., operator location, type of commodity releases) are categorical, numerical data fields have also been categorized into bins to maintain consistency in the data. For example, the age of the item involved in the incident is categorized into 9 bins: 10, 20, 30, 40, 50, 60, 70, 80, > 80. Here, bins 10 and 20 represent ages of the item ≤ 10, and > 10 and ≤ 20, respectively. Some further refining has been done to retain the most informative part of some data fields. For example, the local time of the incident has been extracted as either ’day’ or ’night’ and used for analysis.

The selected data fields, their numbers of categories and the categories are listed in Table 4.1. Among the selected data fields, the data fields such as the type of commodity released, area

of incident, depth of cover, subpart of the system involved, and item involved are selected to infer information of the system that is highly likely to undergo an incident. Further, the equipment specification such as coating type, diameter, and wall thickness of the pipe is selected to give specific information about the pipeline. Here, the pipeline function specifies that the pipeline is either transporting the commodity from the production site/well to refinery or similar facilities (gathering), or from refinery to final use or port (trunkline/transmission). It also indicates if the pipeline is operating above or below the 20 percent of the specified minimum yield strength (≤ 20% SMYS or > 20% SMYS).

Next, among data fields specific to corrosion failure, inspection-related data fields are selected to infer information about the condition of the pipeline. The internal inspection tool indicator represents the pipeline configuration to accommodate internal inspection tools, and the operation complications indicator represents the presence of operational factors which significantly complicate the execution of an internal inspection tool run. Here, SCADA in-place indicator and CPM in-place indicator represent the presence of supervisory control and data acquisition (SCADA)-based system and computational pipeline monitoring (CPM) leak detection system in place on the pipeline or facility involved in the incident, respectively. As a condition monitoring data field, prior damage is selected to represent observable damage to the coating or paint in the vicinity of the corrosion. Data fields such as corrosion inhibitors, corrosion lining, and cleaning dewatering are selected to represent commodity treatment with corrosion inhibitors or biocides, presence of interior coating or lining with a protective coating, and routine utilization of cleaning/dewatering pigs (or other operations).

Table 4.1 Input data fields, their numbers of categories and the categories

Data fields	No. of categories	Categories
Operator location	18	TX, GA, CA, WY, PA, OK, IL, KS, AK, CO, OH, MD, UT, HI, NJ, NY, MT, NH
Local time of incident	2	Day, Night
Type of commodity released	4	Crude oil, Refined and/or petroleum product (non-HVL), HVL or other flammable or toxic fluid, Carbon dioxide/biofuel/alternative fuel
Area of incident	3	Underground, Aboveground, Tank including attached appurtenances/ transitional area

Data fields	No. of categories	Categories
Depth of cover (in)	4	50, 100, 150, >150
System subpart involved	5	Pipeline including valve sites, Terminal/tank farm equipment and piping, Pump/meter station equipment and piping, Breakout tank/storage vessel including attached appurtenances, Equipment and piping associated with belowground storage
Item involved	12	Pipe, Auxiliary piping (e.g. Drain lines), Tank/Vessel, Weld including heat affected zone, Valve, Relief line, Tubing, Meter/Prover, Flange, Scraper/pig trap/Sump/separator, Pump, Other
Part of pipe involved	3	Pipe body, Pipe seam, Others
Diameter of pipe (in)	5	5, 10, 15, 20, >20
Pipe wall thickness (in)	5	0.1, 0.2, 0.3, 0.4, >0.4
Pipeline function	4	> 20% SYMS regulated trunkline/transmission, ≤ 20% SYMS regulated trunkline/transmission, > 20% SYMS regulated gathering, ≤ 20% SYMS regulated gathering
Pipe coating type	11	Coal tar, Fusion bonded epoxy, Cold applied tape, Paint, Asphalt, Extruded polyethylene, Field applied epoxy, Polyolefin, Composite, Others, None
Age of item involved (years)	9	10, 20, 30, 40, 50, 60, 70, 80, >80
Material involved	2	Carbon steel, Others
Internal inspection tools indicator	3	Yes, No, Null
Operation complications indicator	3	Yes, No, Null
SCADA in-place indicator	3	Yes, No, Null
CPM in-place indicator	3	Yes, No, Null

Data fields	No. of categories	Categories
Age of cathodic protection (years)	5	10, 30, 50, 70, >70
Prior damage	3	Yes, No, Null
Corrosion inhibitors	3	Yes, No, Null
Corrosion lining	3	Yes, No, Null
Cleaning dewatering	4	Yes, No, N/A- Not mainline pipe, Null
Age of corrosion inspection (years)	7	1, 2, 3, 4, 5, 6, >6
Age of hydrotest (years)	6	10, 20, 30, 40, 50, >50
Direct inspection type	4	Yes and an investigative dig was conducted at the point of the incident', Yes but the point of the incident was not identified as a dig site, No, Null

As the output of the causal and consequence analysis of corrosion-induced pipeline incidents, four data fields are selected. The output of the causal analysis is the two types of corrosion, i.e., internal and external corrosion. The output of consequence analysis are three data fields: the total cost of property damage (in $s), the net loss of commodity released (in bbls) and the type of release. To increase the computational efficiency of the prediction model, consequences have been categorized into bins of powers of 10 as shown in Table 4.2.

Table 4.2 Output data fields, their numbers of categories and the categories

Data fields	No. of categories	Categories
Cause	2	Internal corrosion, External corrosion
Total cost (in $s)	3	10^5; 10^6; > 10^6
Net loss (in bbls)	3	10^1; 10^2; > 10^2
Type of release	2	Leak, Rupture

4.3.2 Data for probability prediction

To develop the probability prediction model for corrosion-induced pipeline incidents, the number of such incidents in the historical incident database is utilized. The number of such incidents that occurred due to internal and external corrosion over 2010-2019 vary widely as

shown in the 2nd and 3rd columns of Table 4.3, respectively. The total miles of pipeline operations also vary with time, and hence, incident probability prediction using actual numbers of incidents can be grossly misleading. The numbers of incidents are thus normalized by dividing them by the total miles of pipeline operation per 105 miles per year during the time considered.

Table 4.3 Normalized number of corrosion-induced pipeline incidents due to internal and external corrosion that occurred in the USA over the years 2010-2018

Year	Actual number of corrosion-induced pipeline incidents		Total miles of operation in a year	Miles of operation/10^5 miles-year	Normalized number of corrosion-induced pipeline incidents	
	Internal	External			Internal	External
2010	43	28	181986	1.82	23.63	15.38
2011	48	31	183580	1.84	26.15	16.88
2012	66	32	186221	1.86	35.44	17.18
2013	46	27	192412	1.92	23.91	14.03
2014	57	29	199793	1.99	28.53	14.51
2015	52	48	208620	2.09	24.93	23.00
2016	54	36	212109	2.12	25.46	16.97
2017	49	26	215994	2.16	22.69	12.03
2018	31	18	219037	2.19	14.15	8.21

4.4 Proposed Framework

To develop an integrated risk prediction model of corrosion-induced pipeline incidents, a framework that includes ANN models and Bayesian analysis is proposed in this work as shown in Figure 4.1. First, ANN models are utilized to leverage the rich pipeline incident database using data fields selected in Section 4.3.1 and to predict causes and consequences of pipeline incidents. Next, Bayesian analysis is used to determine the probability of pipeline incidents with consideration that the pipeline failure rates vary over time. The information of the predicted cause, consequence, and probability of pipeline incidents are combined to predict the risk of pipeline incidents.

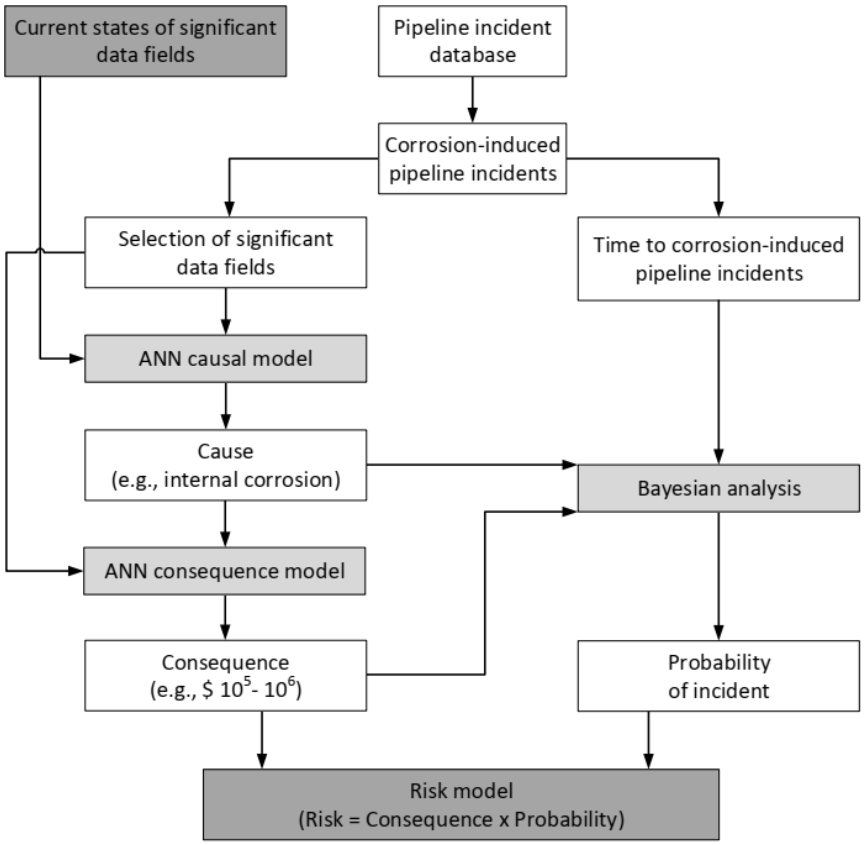

Figure 4.1 The proposed integrated framework for risk prediction of corrosion-induced pipeline incidents

4.4.1 Prediction model for cause and consequences of incident

Utilizing the selected data fields in Section 4.3.1, four ANN models are developed to predict the causes and consequences of corrosion-induced pipeline incidents: one for causal analysis, and three for the consequence analysis in terms of total cost (in $s), net loss (in bbls) and type of release (leak or rupture). Then, the performances of the developed ANN models are evaluated.

ANN model development

The causal and consequence models developed in this work are input-output ANN models with the selected data fields listed in Table 4.1 as their inputs, and the data fields listed in Table 4.2 as their outputs. The developed ANN models capture the causal dependencies and the contribution of the input data fields in the pipeline failures. These models understand the synergy among underlying input data fields and their collective ability to affect the pipeline integrity by utilizing a wealth of empirical knowledge accumulated in the PHMSA database. The methodology followed to develop the ANN models is described in Figure 4.2.

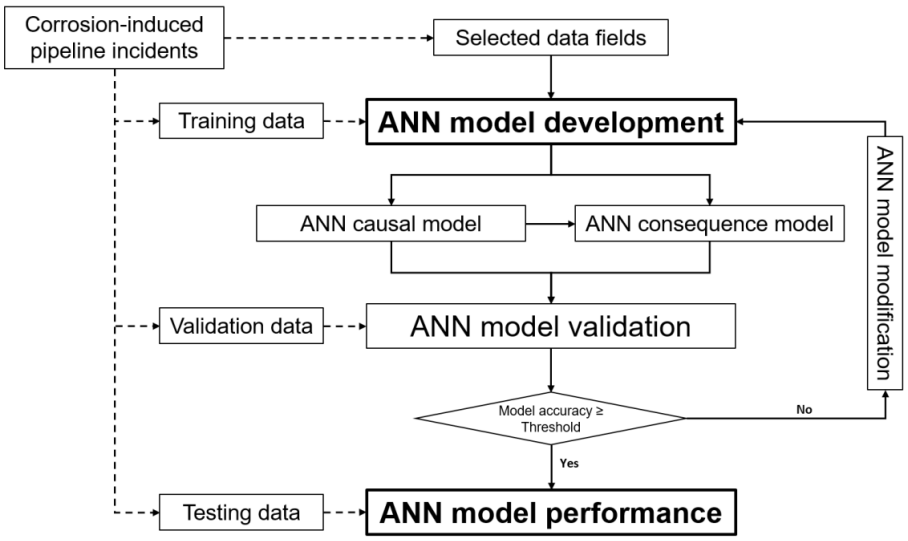

Figure 4.2 ANN model development methodology

To develop the ANN model, first, the contributing factors of a corrosion-induced pipeline incident (i.e., significant data fields) are selected from the PHMSA database as described in Section 4.3.1. Then, the entire corrosion-induced pipeline incident data has been randomly divided into a ratio of 60:20:20 as training, validation, and testing data, respectively. The training and validation data is utilized for ANN model development, while testing data is used for evaluating the performance of the developed ANN models. Specifically, the training data is utilized to fit the ANN model by obtaining its parameter, i.e., weights and biases. Weights represent the strength of connections between neurons of the ANN, and biases are used with inputs to generate the outputs of the ANN. It is to be noted here that since the input and output data fields are categorical, one hot encoding is employed to convert them to numerical values to be utilized with the ANN model. One-hot encoding is a sparse encoding approach that has been widely applied to represent the categorical variables in various fields (Oyedele et al. 2021). In one-hot encoding, each unique value of categorical variables is converted into a new variable with values as 1 and 0 denoting the presence and absence of this new variable (He et al. 2018).

In the ANN model development, the model activation functions have a significant impact on the learning speed of the neural network. Hence, a sigmoid activation function is used for the hidden layers due to its better gradient propagation and efficient computation. Since the outputs of the last layer, i.e, data fields from Table 2, are categorical in nature, a softmax activation function is utilized due to its suitability for a layer with categorical output. Here, the

softmax activation function gives a probability score to each category, and the category with the highest probability score is considered as the output of the network. While learning the model, loss function i.e., the mean square error between the actual output and the predicted output is minimized to obtain the optimized value of weights and biases. Here, a categorical cross-entropy loss function is used for its suitability to categorical outputs.

Next, the validation data is used to provide an unbiased evaluation of model fit on the training data while tuning model hyperparameters, i.e, the number of layers and neurons in each layer. Specifically, the number of layers and neurons in each layer are adjusted to obtain a model accuracy greater than a predefined threshold.

ANN model performance evaluation

To evaluate the performance of the developed ANN model, the testing data, which is not presented to the network during model development, is utilized. Specifically, using the developed ANN models, output for each point in the testing data is predicted. The model accuracy is calculated as follows:

$$Model\ accuracy = 100 * \frac{\sum_{k-1}^{C} p_k}{N} \quad (1)$$

where p_k is the number of accurately predicted output in k^{th} output category, C is the total number of output categories, and N is the total number of data in the testing data. Utilizing the developed ANN causal model, the cause of a corrosion-induced pipeline incident (i.e., internal or external corrosion) is predicted given the current condition of the pipeline. Next, to predict the consequence of a corrosion-induced pipeline incident, the ANN consequence models are developed in the same manner. It is to be noted that the ANN consequence models utilize the predicted cause of the pipeline incident as one of the inputs in addition to the selected data fields. Next, the probability of the corrosion-induced pipeline incident is estimated using Bayesian analysis.

4.4.2. Prediction model for probability of incident

Prediction of the probability of an incident over a given time requires an understanding of the current incident rate and trend. From Table 4.3, it can be seen that the number of incidents varies over the years, indicating that the times between incidents are not uniformly distributed. In other words, incidents seem to occur more frequently in some years than others. When data is cluttered in certain periods and dispersed over others, it is not possible to provide an accurate

prediction into the future using average values, i.e., by simply assuming that the incident rate does not change with time. A time-trend analysis of incidents is required to consider whether the incident rate is increasing, decreasing, or remaining constant over time.
The Poisson process commonly defines the failure of an entity by taking a failure process as being points located randomly in the time-space, thus enabling the Poisson process to analyze time-series data. Relaxation of a constant incident rate leads to consideration of a nonhomogeneous Poisson process (NHPP), where the incident/failure rate, called the rate of occurrence of failure (ROCOF) is a function of time, λ(t). For systems whose failures (or incidents) are influenced by multiple contributors or causation mechanisms, the NHPP is particularly suitable (Rausand and Hoyland 2003). In this work, we use the assumption that failures follow a NHPP to predict the time to the next failure and its probability.

Nonhomogeneous Poisson process

In NHPP, the incidents do not require stationary increments, i.e., some incidents are more likely to occur at certain times than others, and the time between incidents are generally neither independent nor identically distributed (Rausand and Hoyland 2003). Among the various parametric models that can define a NHPP, the power law model is the most developed and had been used in this study for trend analysis. In the power law model, the ROCOF is defined as

$$\lambda(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \qquad \alpha, \beta \geq 0 \tag{2}$$

where α is the scale parameter that sets the units with which time is measured, and β is the shape parameter that determines how the ROCOF changes over time. The use of power law gives the benefit of direct inference about the trend of failures/accidents through the parameter β. Value of $\beta < 1$ indicates a decreasing trend in the rate of incidents over time, while $\beta > 1$ indicates incidents are occurring more frequently as time increases. Thus, a benefit of modeling a system as a NHPP is that it allows monitoring of the system's performance over time. The parameters α and β will have to be estimated to determine the ROCOF at a given time t. The current study uses Bayesian inference to determine these parameters from the incident database. The reason for the preference of using Bayesian analysis over other methods, such as the maximum likelihood estimate, lies in the data being processed. Incidents of pipeline failure are taken as random processes, influenced by multiple causes and contributors, and Bayesian analysis assumes the value for the time to next failure (the parameter of interest) lies within a fixed credible range. Frequentist methods, unlike Bayesian analysis, assume this parameter of interest to have a fixed

value and hence are not appropriate for modeling such random pipeline incidents (Halim et al. 2020).

Bayesian inference as a tool of prediction

According to the Bayesian theorem, the posterior distribution of the parameter of interest is written as:

$$\pi_1(\theta|x) = \frac{f(x|\theta)\pi_0(\theta)}{\int_\theta f(x|\theta)\pi_0(\theta)\,d\theta} \tag{3}$$

where θ is the parameter of interest, $\pi_0(\theta)$ is the prior distribution of θ, f(x|θ) is the likelihood function, or the aleatory model of x given values of θ, and π1(θ|x) is the posterior distribution of θ. In this work, the parameters of interest are those that enable the determination of the incident rate (α and β for power law), and x is the observed data.

Bayesian analysis using NHPP is implemented in OpenBUGS R using the algorithm described in (Kelly 2007, Halim et al. 2021). This algorithm uses Markov chain Monte Carlo (MCMC) sampling to generate the joint posterior distribution of α and β. Observed data are incorporated into the model through the likelihood function. For the NHPP, each incident time after the first is taken to depend on the preceding incident time (Rodionov et al. 2009). If ti is the cumulative operational time incurred from the first incident to the i^{th} one, the likelihood function for the power law process is given by:

$$f(t_1, t_2, \cdots t_n|\alpha, \beta) = \frac{\beta^n}{\alpha^{n\beta}} \prod_{i=1}^{n} t_i^{\beta-1} exp\left[-\left(\frac{t_i}{\alpha}\right)^\beta\right] \tag{4}$$

To make the analysis completely dependent on the observed data, non-informative priors are chosen for the Bayesian analysis. In OpenBUGS R , these are inserted as:

$$\alpha \sim gamma(0{:}0001; 0{:}0001) \tag{5a}$$

$$\beta \sim gamma(0{:}0001; 0{:}0001) \tag{5(b)}$$

The algorithm developed in OpenBUGS R uses the observed data to update the non-informative prior and obtain the joint posterior distribution of α and β. These are then used to predict the next value (with credible interval) of the cumulative time to failure. Using this, the probability of an incident over a given time in the future is calculated using:

$$Pr(T \leq t) = F(t) = 1 - exp\left(-\int_T^{T+t} \mu(\tau)\,d\tau\right) = 1 - exp\left(-\mu\left[(T+t)^\beta - T^\beta\right]\right) \tag{6}$$

T +t where $\mu = \alpha - \beta$, T is the cumulative operating time upto the last incident, and t is the additional time of operation over which the probability of a pipeline incident is being predicted. For the iterative MCMC simulation, two chains are run simultaneously to determine quantitatively when convergence is achieved between the two chains using the BGR (BrooksGelman-Rubin) diagnostic in the OpenBUGS R software (Rodionov et al. 2009). Once convergence is achieved, the iterations are discarded and the model is re-run for further simulations that estimate the parameters.

Utilizing the above-described models, first, the cause of a corrosion-induced pipeline incident (i.e., internal or external corrosion) is predicted given the current conditions of a pipeline. Next, the consequence and probability of pipeline failure due to the predicted cause are estimated. They are then multiplied to predict the risk of a pipeline failure due to the predicted cause of a corrosion pipeline incident. In the following sections, the results obtained from modeling the ANN and Bayesian analysis with NHPP are detailed and discussed.

4.5 Results and Discussion

The proposed framework including ANN models and Bayesian analysis is developed utilizing the preprocessed data for corrosion-induced pipeline incidents. The performance of the developed models is demonstrated on testing data.

4.5.1 Cause and consequence prediction using ANN models

Using 26 inputs listed in Table 4.1, four ANN models are constructed: one for causal analysis and the other three for consequence analysis. The ANN causal model differentiates between causes of incident as internal corrosion and external corrosion. The ANN consequence models predict the consequence of the incident in terms of the total cost of property damage (in $s), the net loss of commodity released (in bbls), and the type of incident. These four ANN models are validated and tested for their prediction accuracy.

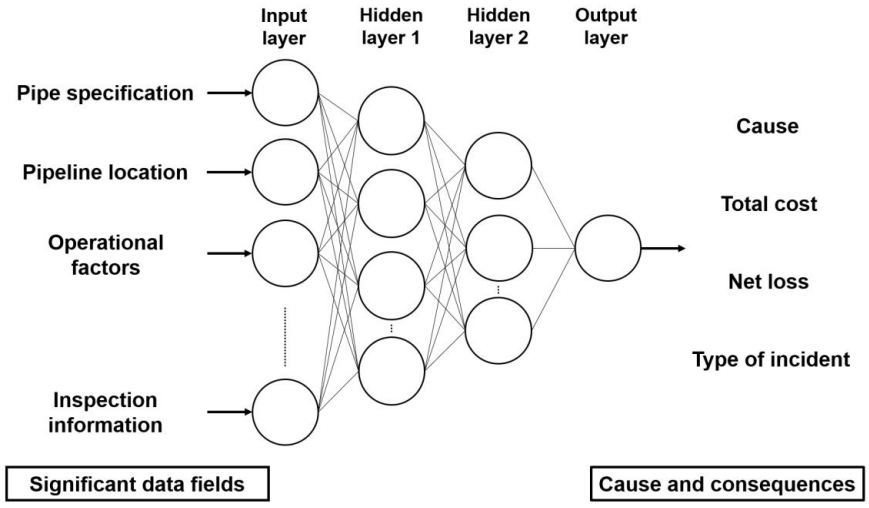

Figure 4.3 Structure of ANN model

The final structure of the ANN model developed in this work after training and validation is presented in Figure 4.3. Here, the model accuracy threshold for model development is taken as 90%. The resultant network structure is designed to have an input layer, two hidden layers, and an output layer. Here, the inputs of the model are the selected data fields listed in Table 4.1, which are connected to the neurons/nodes in the first hidden layer. The first hidden layer is designed to have 20 nodes, and they are connected to the second layer which has 25 nodes. The second hidden layer is connected to the output layer, i.e., the data fields listed in Table 4.2. For each data field listed in Table 4.2, an ANN model is developed.

The performance of the four developed ANN models on the training data are compared using the learning rate parameter which determines the rate to move toward a minimum of a loss function at each iteration. It can be observed in Figure 4.4 (left) that the learning rate of ANN models with outputs as cause and release type are higher than that of ANN models with outputs as net loss and total cost are low. This can be explained on the basis of the number of output categories of ANN models. Learning of causal dependencies among the input data fields to predict a higher number of output categories is difficult. Therefore, the learning rate is lower for ANN models with outputs as net loss and total cost. Since a lower learning rate implies lower model accuracy, the accuracies of ANN models with outputs as net loss and total cost are lower than that of ANN models with outputs as cause and release type as shown in Figure 4.4 (right).

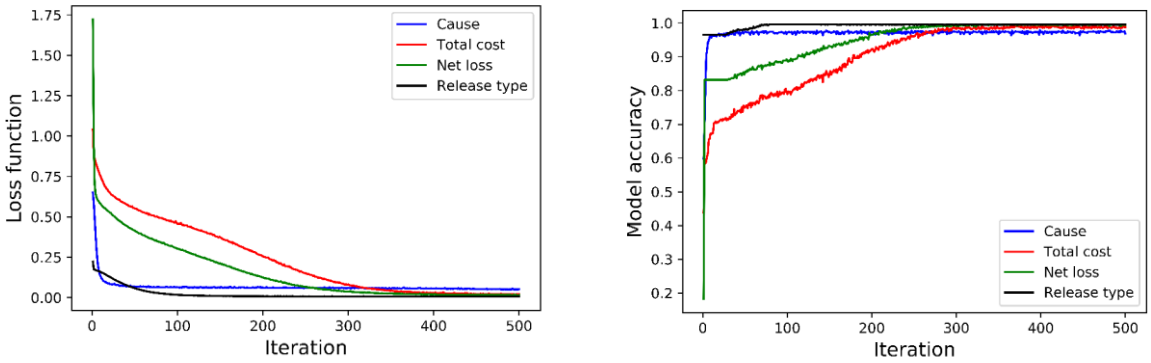

Figure 4.4 Loss function vs iteration (left) and model accuracy vs iteration (right) for training data

The performances of the developed ANN models are calculated using (1), where N = 144 and the number of output categories of each model, C, is listed in the 2nd column of Table 4.4. The model performances on the validation data are listed in the 3rd column of Table 4.4. The model accuracy of all of the models on the validation data is greater than the threshold accuracy, i.e., 90%. Next, the model performances of the developed ANN models are evaluated on the testing data and listed in the 4th column of Table 4.4. It can be seen that model accuracy decreases with an increase in the number of categories. Another reason for a lower model accuracy for total cost is due to fact that total cost of an incident is affected by other factors such as population and natural resources near the pipeline, and presence of ignition source, which are not present in the PHMSA database.

Table 4.4 ANN model accuracy: Validation and testing

Output	No. of categories	Validation	Testing
Cause (Internal/External corrosion)	2	97.40	94.54
Total cost (10^5; 10^6; > 10^6 in $s)	3	90.80	70.08
Net loss (10^1; 10^2; > 10^2 in bbls)	3	95.76	79.31
Type of release (Leak/Rupture)	2	94.53	91.72

After predicting the cause and consequence of a corrosion-induced pipeline incident given current conditions of a pipeline, the probability of incident due to the predicted cause is estimated using Bayesian analysis with NHPP. It is to be noted that among the predicted consequences, i.e.,

the total cost, net loss and type of release, only the total cost is used for further analysis since it provides the most quantitative representation of consequence of an incident.

4.5.2. Probability prediction using Bayesian analysis with NHPP

To predict the probability of a corrosion-induced pipeline incident, the incidents are divided into 6 categories according to the cause and the total cost of the incident (Table 4.5). The reason for this categorization is that the probabilities of corrosion-induced pipeline incidents vary significantly for different causes and total cost of incidents. As seen from Table 4.5, the number of pipeline incidents due to internal corrosion are higher than that due to external corrosion, and low consequence incidents are in higher numbers than those with high consequences, as expected. Due to variation in the number of incidents for different causes and total cost of incidents, the probabilities of the incidents may also vary.

Table 4.5 Number of incidents in the categories based on causes and costs of incidents (TC = Total cost in $)

	Number of internal corrosion-induced pipeline incidents			Number of external corrosion-induced pipeline incidents		
Year	$TC \leq 10^5$	$10^5 < TC \leq 10^6$	$TC > 10^6$	$TC \leq 10^5$	$10^5 < TC \leq 10^6$	$TC > 10^6$
2010	33	10	0	17	9	2
2011	34	12	2	14	12	5
2012	46	12	8	22	8	2
2013	23	16	7	15	8	4
2014	33	18	6	15	12	2
2015	32	19	1	30	15	3
2016	37	15	2	17	16	3
2017	25	20	4	14	7	5
2018	17	10	4	9	6	3

For each of the category of corrosion-induced pipeline incidents, a Bayesian analysis model is developed as described in Section 3.2.2 to predict the probability of incidents. The cumulative time to incident is calculated from 01/01/2010 up to the incident of interest, divided by the normalizing factor (miles of operation per 105 miles per year). The observed data is provided as supplemental information. The developed algorithm uses the observed data to update the parameters α and β. The values of these parameters are given in Table 6. Since $\beta < 1$ and $\beta >$

1 indicate a decreasing and increasing rate of incidents with time, the internal corrosion-induced pipeline incidents with a total cost ≤ $10^5 are observed to have a decreasing rate of incidents with time, while all other categories are seen to have an increasing trend. It can also be observed from the increasing value of β with the increase in the total cost of the incident that the number of incidents are increasing at a higher rate for the incidents with higher total cost for both of the internal and external corrosion-induced incidents.

Table 4.6 Bayesian parameters: α and β (TC = Total cost (in $s))

		α		B	
Cause	Total cost	Mean	Std. dev.	Mean	Std. dev.
Internal corrosion	$TC \leq 10^5$	5.53	0.07	0.986	0.0024
	$10^5 < TC \leq 10^6$	35.83	12.22	1.266	0.1121
	$TC > 10^6$	121.70	56.29	1.346	0.2449
External corrosion	$TC \leq 10^5$	17.35	6.49	1.096	0.0882
	$10^5 < TC \leq 10^6$	30.75	12.84	1.127	0.1184
	$TC > 10^6$	118.30	58.47	1.271	0.2434

Utilizing the obtained posterior distributions of α and β ((3)), α and β are sampled and substituted in (6) to obtain the probability of the incident in next 7 days. In (6), t is the additional time of operation over which the probability of a corrosion-induced pipeline incident is being estimated, i.e., 7 days. The mean and standard deviation of the predicted probability of pipeline incident is presented in the 3rd and 4th columns of Table 4.7, respectively. It can be observed that the probability of a pipeline incident with a higher consequence is significantly lower than that with a lower consequence. This observation is in agreement with the observation from historical data presented in Table 4.5 in which the frequency of a pipeline incident with a higher consequence is significantly lower than that with a lower consequence.

To establish the credibility of the developed Bayesian analysis models, the calculated next time to the incident data is compared with the historical data in each category as described in (Halim et al. 2021). The mean and standard deviation of the predicted next time to the incident are reported in the 5th and 6th columns of Table 4.7, respectively. It can be observed that the actual next time to the incident for each category is within one standard deviation from the predicted

mean next time to the incident. Therefore, the developed Bayesian analysis models can be utilized to accurately predict the probability of pipeline incident.

Table 4.7 Bayesian analysis model accuracy: Predicted probability of incident, and predicted and actual next time to incident (TC = Total cost (in $s))

		Probability of incident		Next time to incident (days since 01/01/2010)		
Cause	Total Cost	Mean	Std. dev.	Mean	Std. dev.	Actual
Internal corrosion	$TC \leq 10^5$	0.6942	0.03024	1615	5.85	1621
	$10^5 < TC \leq 10^6$	0.5066	0.04386	1620	9.96	1622
	$TC > 10^6$	0.1771	0.04046	1568	39.30	1587
External corrosion	$TC < 10^5$	0.5111	0.0405	1617	10.04	1616
	$10^5 < TC < 10^6$	0.3597	0.04259	1633	16.12	1624
	$TC > 10^6$	0.1478	0.03747	1589	48.69	1615

Utilizing the predicted cause and consequence of a corrosion-induced pipeline incident given the current conditions of a pipeline using ANN models, one of the six Bayesian analysis models is selected to predict the probability of pipeline incident in the next 7 days. Then, the predicted consequence and probability of pipeline incident are multiplied to predict the risk of a pipeline incident in the next 7 days (as described in Figure 6). As an example, to illustrate the risk prediction, the current conditions of the pipeline corresponding to the pipeline incident occurred on 12/21/2018 in Houston, TX (PHMSA report key 20190015) is utilized. The true cause, total cost, net loss, and type of this incident are internal corrosion, $16778, 0 bbl, and leak, respectively. Utilizing the developed ANN models with the current condition of the pipeline, the cause, total cost, net loss, and type of this incident are predicted as internal corrosion, $\leq \$10^5$, $\leq$ 10 bbl, and leak, respectively. An overestimation of the total cost and net loss occurs while using ANN models. Such an overestimation can be attributed to the categorization of these model outputs into bins and can be reduced by considering a higher number of bins for model outputs. However, it will result in a higher number of output categories, and consequently, leads to a lower model accuracy (as discussed in Section 4.5.1). Therefore, the number of bins for model outputs should be selected carefully to balance the trade-off between model output overestimation and the model accuracy.

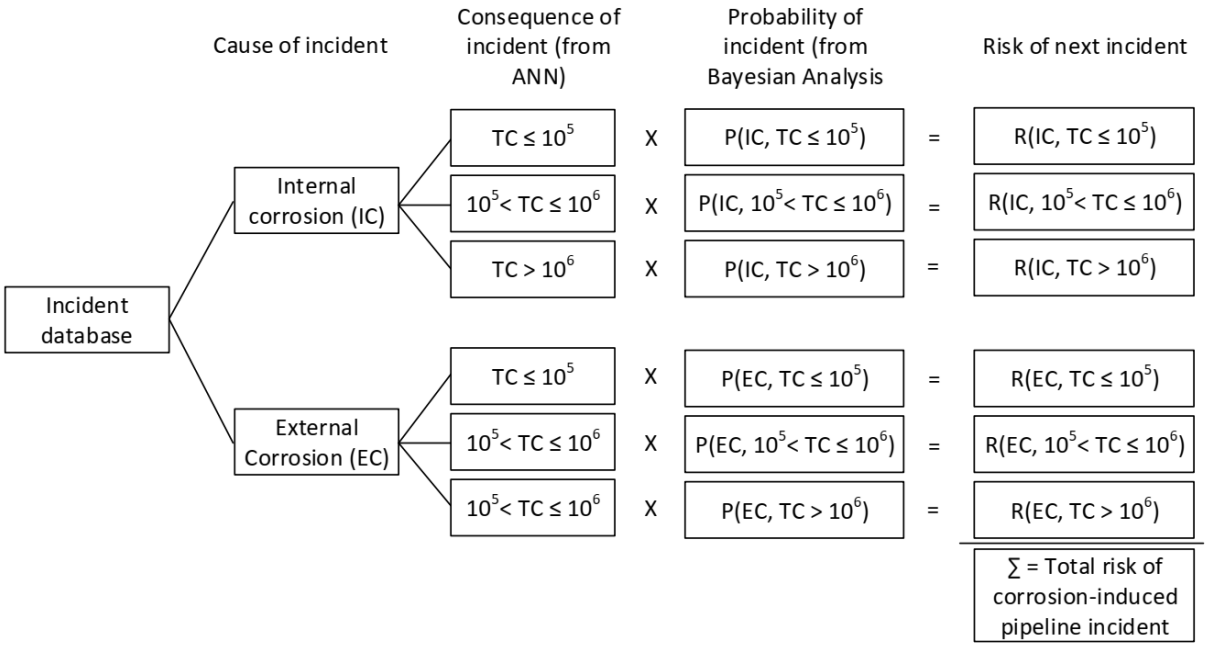

Figure 4.5 Risk calculation framework

Further, based on the predicted cause and consequence, i.e., internal corrosion and ≤ $\$10^5$, the first Bayesian analysis model is selected to predict the probability of such incident in the next 7 days, i.e., P (IC, TC ≤ 10^5). As seen from Table 4.7, the mean and standard deviation of P(IC, TC≤10^5) is 0.6942 and 0.03024. Multiplying the consequence and probability of the incidence, the mean of R (IC, TC ≤ 10^5) is calculated as ≤ $\$6.942 \times 10^4$. Considering one standard deviation above the mean as its spread, the upper limit of R (IC, TC ≤ 10^5) is calculated as $\$7.244 \times 10^4$ (= (0.6942 + 0.03024) x $\$10^5$). It is to be noted that rest of the risks shown in Figure 4.5 are equal to zero since they are associated with a combination of causes and consequences other that the ones predicted. Hence, the upper limit of risk given current conditions of the pipeline as for the incident with the PHMSA report key 20190015 is predicted as $\$7.244 \times 10^4$. Here, the predicted risk value of a pipeline given its current condition provides a valuable insight into the impending loss and can be utilized by the pipeline operators for strategic management of the impending loss by controlling the factors that may result into the predicted cause of the incident.

5.0 CONCLUSION

In this study, pipeline incident data obtained from incident records and incident investigation reports has been analyzed. The report examined the available sources of incident data and carefully selected a few datasets for further analysis. A few NLP and text mining techniques have been explored to extract useful causal information from incident data. Later an ANN model has been developed to predict pipeline incidents for specific conditions.

The current work analyzes three incident datasets collected from pipeline operations from three different regions: PHMSA from the USA, NEB from Canada and EGIG from European Union. All databases provide a large amount of information related to an even larger number of pipeline incidents. A review of the type of information gathered is provided and some analysis of data pertaining to the causal factors behind the failures are provided. PHMSA database provides information on background factors, while NEB dataset provides information on underlying causes and allows identification of multiple causes behind an incident. Management issues influencing pipeline operations are brought about by the underlying factors in the NEB database.

Although several studies investigated the causal relationship amongst different factors leading to pipeline failure, the relationships were merely associative in nature. It suggests that establishing a cause-effect relationship from such data is difficult. Pipeline specific information and operational data (coined as background factors) are essential, but they are not adequate to provide an in-depth understanding of the pipeline failures. Hence, the scope of improvement studying past incident data is limited. It requires “true” roots causes, or in other words, the underlying causes (such as management or organizational causes). For instance, association of crude oil or natural gas with a failure or association of pump or tank with a corrosion failure is not sufficient. More causal information such as incorrect operation or maintenance related to corrosion are required. Because crude oil or pump cannot be attributed as the cause of incident, however, incorrect operation or inadequate maintenance can be. NEB dataset provided a very good classification of such underlying causes. Such data will guide the development of a causal model that can capture the causation and interdependence of the factors under consideration.

A few NLP and text mining techniques have been explored to extract useful information that may provide some incite about the underlying causes of the pipeline incidents. Current study explored the capability of three techniques (K-means clustering, co-occurrence network, and topic modeling) extracting contributing factors and causality of incidents from both the narrative

comments and incident investigation reports. It introduced a workflow of automated content analysis of incident descriptions. Due to the vast amount of text data in safety area and due to the lack a generalize NLP technique, no existing package is currently in place to offer automated solutions to extract causations and identify contributing factors. The workflow therefore provides a novel solution to mine the hidden knowledge to enhance the learning from the past incidents. The workflow can produce a cause-and-effect storyline of incident and the information unlocked allows a quick understanding of knowledge accumulated in the incident narratives which is time-consuming and labor-intensive to digest manually even by experts. Thus, this work can provide a potential improvement of automation to previous works on risk analysis and accident modeling.

Co-occurrence network approach shows high potential in turning incident text data into valuable knowledge because the structure of network may lead to a hierarchy of causation. By applying the co-occurrence network to narrowed-down narrative datasets under specific cause labels, more fine-grained factors of incident are identified which demonstrates the scalability of co-occurrence network approach. Topic modeling technique was also found to be promising capable to identifying factors relevant to a specific cause. Since topic modeling has the capability to operate as a semi-supervised learning method, it can be trimmed for specific extraction requirement. Despite the results by K-means clustering being coarse compared to the other, the case shown in the work still demonstrates that clustering analysis can be an asset to uncovering the synergistic effect of causes. t-SNE, is found to be promising in dealing with significantly high dimensional TF-IDF matrix as dimensionality reduction technique combined with clustering algorithm.

At the end, an integrated framework for risk prediction has been presented. The framework was applied to a corrosion-induced pipeline incidents in the onshore HL transmission pipelines, enabling determination of the causes of incidents, their subsequent consequences and probabilities. The causal and consequence estimation models have been developed utilizing the ANN technique, and the probability estimation model utilizes the Bayesian analysis. The proposed framework first collects and processes the onshore HL transmission pipeline incident data from the PHMSA database. It eliminates the redundant data fields and selects 70 data fields resulting in higher information content. The number of data fields is further reduced to 26 using process knowledge resulting in higher information density. Utilizing the selected data fields, a reasonably accurate prediction model is developed to predict the cause, consequence, and probability of corrosion-

induced pipeline incidents. Particularly, the proposed ANN model is applied to the preprocessed incident data and validated with 90-95% accuracy. The Bayesian analysis model performance is also tested for the prediction of the probability of the incident. Utilizing the proposed framework including ANN models for cause and consequence prediction and Bayesian analysis for probability prediction, the risk of a corrosion-induced pipeline incident can be predicted given the current condition of pipelines. This shows the strength of the proposed framework to predict the risk of corrosion-induced pipeline incidents and can further be extended to pipeline incidents caused by other causes such as excavation, natural forces, etc.

6.0 FUTURE WORK

The current study examined the available incident data and incident investigation reports and investigated the scope of developing a causal model and predicting future pipeline incidents. The study explored different techniques and developed framework and models to extract information to build the causal model. Although the incident data collection method is very robust and comprehensive, there are opportunities to improve. Some important data field can be added and/or refined. A summary of scopes of improvement has been given below:

- Data and information regarding all three aspects of causal analysis need to be considered. Currently, PHMSA is gathering direct causes (mapped cause and sub-causes), background factors (pipeline specific information and operational data) and a little bit of underlying factors (such as maintenance/ inspection data). Although the incident reporting system is very comprehensive, the underlying factors gathered are inadequate to understand the true condition of the pipeline and how to improve. A set of underlying conditions has been defined by NEB. Similar causes or reasoning can be adopted by PHMSA and accommodate them in incident reporting system.
- Currently PHMSA allows to report only one mapped cause. However, there were incidents where two simultaneous causes contributed to the incident. Reporting of multiple causes should be allowed.
- All data collected through the incident recording systems must have end purposes. Some specific purposes are certainly evident; however, any overarching goal is absent. In the absence of true underlying causes, the generation of effective actionable learning is difficult from the current dataset. For instance, data tells us about that equipment failure is a major cause of pipeline incident. It also gives us a clear picture of the corresponding consequences. However, how to overcome the problem, whether is it due to lack of maintenance or poorly designed system, cannot be identified from it. Such objectives need to be defined and any data necessary to complete the cycle must be included in the incident data recording system.
- Incident investigation and findings can be a very important tool for learning from past. However, to make it a successful process, the incident investigation process needs to be designed appropriately. Currently, there is no common structure or methodology for incident investigation process. It is apparent from the most incident investigations had a

specific purpose that is to identify the precise direct cause which the reports termed as root cause to identify. Mostly processes did not intend to uncover the underlying causes of the incidents. Incident investigation process must identify all possible underlying causes in addition to the specific purpose they are currently being employed for.

- A general classification of causes or contributing factors (taxonomy) has been provided here. However, a more comprehensive study is needed to define the terminologies precisely. Existing human and organizational factors classification can be adopted, or a new system can be developed.
- Incident investigation reporting should be more structured. There must be a few mandatory sections such as incident description, key finding and/or recommendations or key lessons. NLP or text mining techniques will be able to extract necessary information from these sections more effectively.
- More NLP and/or text mining techniques should be explored to improve their effectiveness. With better sets the techniques will be more efficient.
- Although ANN is a powerful tool, it has its limitations. In current context, the lack of data or inadequate data forces to ignore the data field from the ANN model even though they seem reasonably relevant and important. Bayesian network with support from expert elicitation can be useful for such scenarios. Bayesian network approaches should be parallelly explored.

APPENDIX

Appendix A1 A list of incident investigation reports mentioned in PHMSA website

Operator Name	Operator Id	System Type	Apparent Cause	State	Failure Date	Date Posted
Magellan Pipeline Company, LP	22610	HL	Excavation Damage 3rd Party	MN	4-Nov-03	12-Jun-12
Enbridge Energy LP	11169	HL	Natural Force	MN	19-Feb-04	17-Oct-10
BP Pipeline Co	18386	HL	Corrosion External	OH	25-Feb-04	26-Dec-12
Jayhawk Pipeline	9175	HL	Corrosion Internal	KS	12-Apr-05	13-Oct-11
Southern Star Central Gas Pipeline, Inc.	31711	NG	Incorrect Operation	KS	30-Jun-05	26-Dec-12
Amoco Oil Company	395	HL	Equipment Failure	IN	18-Aug-05	20-Feb-13
TE Products Pipeline Co	19237	HL	Incorrect Operation	OH	18-Sep-05	14-Feb-12
Enbridge Energy LP	11169	HL	Weld Seam Failure	WI	1-Jan-07	11-Feb-13
Southern Natural Gas	18516	GT	Material Failure - Weld	AL	23-Jan-07	7-Apr-11
Enbridge Energy LP	11169	HL	Weld Failure	MN	13-Nov-07	17-Nov-10
Northern Natural Gas Co	13750	GT	Natural Forces Damage	MN	23-May-08	8-Apr-16
Panhandle Eastern Pipeline Co	15105	GT	Corrosion External	MO	25-Aug-08	25-Oct-10
Marathon Pipe Line, LLC.	32147	HL	Material Failure Rupture	IL	3-Sep-08	6-Jun-17
Dominion Transmission Inc	2714	GT	Excavation Damage 2nd Party	WV	23-Oct-08	29-Nov-10
Columbia Gas Transmission	2616	GT	Material Failure - Cracking	PA	5-Nov-08	22-Mar-11
Columbia Gas Transmission	2616	GT	Corrosion Internal	WV	4-Jan-09	20-Sep-11
Mid-Valley Pipeline Co	12470	HL	Material Failure	OH	18-Feb-09	17-Nov-10
Hampshire Gas	7050	GT	Other - Miscellaneous	WV	24-Mar-09	9-Mar-11
Columbia Gas Transmission	2616	GT	Equipment Failure	WV	4-Apr-09	22-Mar-11
Columbia Gas Transmission	2616	GT	Equipment Failure	WV	19-Apr-09	22-Mar-11

Operator Name	Operator Id	System Type	Apparent Cause	State	Failure Date	Date Posted
Columbia Gas Transmission	2616	GT	Equipment Failure	WV	7-May-09	22-Mar-11
Enbridge Energy LP	11169	HL	Incorrect Operation	WI	21-May-09	11-Feb-13
Enbridge Energy LP	11169	HL	Material Failure	MN	9-Jun-09	17-Oct-10
Enterprise Products Operating LLC	31618	GT	Incorrect Operation	OCS	4-Aug-09	13-Jan-12
Texas Gas Transmission LLC	19270	GT	Corrosion Internal	TX	4-Aug-09	9-Jun-11
Explorer Pipeline Co	4805	HL	Corrosion External	OK	17-Aug-09	10-Nov-10
El Paso Natural Gas	4280	GT	Unknown, Miscellaneous	TX	5-Nov-09	18-Nov-11
Columbia Gas Transmission	2616	GT	Material Failure - Valve	WV	16-Nov-09	20-Sep-11
National Fuel Gas	13063	GT	Corrosion External	NY	21-Dec-09	22-Mar-11
Enterprise Operating Products	31618	HL	Material Failure - Fitting	TX	23-Dec-09	17-Dec-10
Buckeye Partners LP	1845	HL	Corrosion External	PA	29-Dec-09	4-Apr-11
Southern Natural Gas	18516	GT	Material Failure Pipe	AL	31-Dec-09	9-Jun-11
Southern Natural Gas	18516	GT	Corrosion External	MS	6-Jan-10	10-Aug-11
Enbridge Energy, LP	11169	HL	Material Failure Pipe	ND	8-Jan-10	8-Jan-10
El Paso Natural Gas Company	4280	GT	Material Failure Pipe	AZ	1-Mar-10	9-Jun-11
Mid-Valley Pipeline Co	12470	HL	Corrosion Internal	TX	1-Mar-10	10-Aug-11
Southern Star Central Gas Pipeline	31711	GT	Material Failure Weld	KS	2-Mar-10	15-Feb-12
KM Interstate Gas Transmission Co	1007	GT	Material Failure Pipe	NE	9-Mar-10	17-Jan-12
SFPP LP	18092	HL	Corrosion Internal	CA	16-Mar-10	20-Apr-11
Sunoco Inc R&M	18779	HL	Equipment Failure	PA	25-Mar-10	9-Jun-11
Whitecap Pipe Line Company	31563	HL	Other Outside Force Damage	OCS	25-Mar-10	24-Aug-11
Bridger Lake, LLC	32483	HL	Incorrect Operation	WY	2-Apr-10	24-Aug-11

Operator Name	Operator Id	System Type	Apparent Cause	State	Failure Date	Date Posted
TE Products Pipeline Company, LLC (TEPPCO)	19237	GT	Incorrect Operation	IN	13-Apr-10	10-Jan-17
Williams Gas Pipeline – Transco	19570	GT	Corrosion External	TX	26-Apr-10	20-Apr-11
Chevron Pipe Line Company	2731	HL	Outside Force Damage	UT	11-Jun-10	9-Jun-11
Suncor Energy (Rausand and Hoyland) Pipeline Company	31822	HL	Incorrect Operation	WY	14-Jun-10	17-May-12
Dixie Pipeline Company	3445	HL	Excavation Damage 3rd Party	GA	5-Jul-10	30-Sep-11
Magellan Ammonia Pipeline	12105	HL	Material Failure - Pipe	NE	23-Jul-10	10-Aug-11
Enbridge Energy, LP	11169	HL	Equipment Failure	MN	28-Jul-10	4-Jan-17
Northern Natural Gas Co.	13750	GT	Other Outside Force Damage	IA	6-Aug-10	11-Jan-17
Enterprise Products Operating LLC	31618	HL	Material Failure Pipe	NY	27-Aug-10	13-Jan-12
Harbor Pipeline Co	7063	HL	Incorrect Operation	NJ	11-Oct-10	26-Dec-12
Shell Pipeline Company, LP	31174	HL	Material Failure Pipe	LA	16-Nov-10	26-Dec-12
Tennessee Gas Pipeline Co	19160	GT	Material Failure Pipe	LA	30-Nov-10	13-Jan-12
Chevron Pipe Line Company	2731	HL	Incorrect Operation	UT	1-Dec-10	7-Jul-11
Tennessee Gas Pipeline	19160	GT	Corrosion Internal	TX	8-Dec-10	20-Sep-11
Columbia Gas Transmission Corp.	2616	GT	Weld Leak	NY	11-Jan-11	28-Feb-13
Columbia Gas Transmission Corp.	2616	GT	Weld Leak	NY	11-Jan-11	13-May-16
Chevron Pipe Line Company	18124	HL	Excavation Damage	LA	26-Jan-11	10-Jul-12
Denbury Gulf Coast Pipelines	32545	HL	Material Failure Pipe/ Weld	LA	14-Feb-11	22-Jun-17
Gulf South Pipeline Company, LP	31728	NG	Material Failure	TX	14-Feb-11	18-Nov-13
Enterprise Crude Pipeline LLC	30829	HL	Incorrect Operation	OK	21-Feb-11	26-Dec-12

Operator Name	Operator Id	System Type	Apparent Cause	State	Failure Date	Date Posted
Tennessee Gas Pipeline Company	19160	GT	Material Failure	OH	1-Mar-11	11-Jan-17
Buckeye Partners	1845	HL	Corrosion External	PA	20-Mar-11	20-Sep-11
ExxonMobil Pipeline Company	4906	HL	Outside Force Damage	MT	1-Jul-11	20-Feb-13
TransCanada Northern Border Inc.	32487	GT	Construction Damage	WY	20-Jul-11	26-Feb-13
Central Florida Pipeline Corporation	2190	HL	3rd Party Excavation	FL	22-Jul-11	20-Feb-13
Sunoco Pipeline L.P.	18718	HL	Corrosion Internal	PA	2-Aug-11	22-Mar-16
Chevron Pipe Line Company	2731	HL	Material Failure - Weld	TX	8-Sep-11	12-Jun-12
Buckeye Partners LP	1845	HL	Excavation Damage 3rd Party	NY	20-Sep-11	17-May-12
Magellan Pipeline Company, LP	22610	HL	Third Party Excavation	KS	6-Oct-11	26-Feb-13
Columbia Gas Transmission Corp	2616	GT	Corrosion Internal	PA	3-Nov-11	24-Apr-12
Belle Fourche Pipeline Company	1248	HL Apparent	Operator Error	WY	13-Nov-11	19-Apr-17
Tennessee Gas Pipeline Company	19160	NG Apparent	Material Failure	MS	22-Nov-11	21-Aug-13
Magellan Pipeline Company	22610	HL	Incorrect Operations	TX	1-Dec-11	11-Feb-13
Enterprise Products Operating LLC	31618	HL	Material Weld Failure	TX	27-Dec-11	8-Apr-16
Columbia Gulf Transmission	2620	GT	Natural Forces Damage	KY	1-Feb-12	22-Mar-16
Williams Gas Pipeline-Transco	19570	GT	Corrosion External	NJ	2-Apr-12	22-Mar-16
Enterprise Crude Pipeline, LLC	30829	HL	Internal Corrosion	OK	8-Apr-12	3-Feb-14
Texas Eastern Transmission L.P.'s	19235	GT	Equipment failure - Start Air Valve Malfunction	PA	13-Apr-12	10-Mar-15
El Paso Natural Gas Company	4280	GT	Equipment Failure	CA	2-May-12	23-Jul-13

Operator Name	Operator Id	System Type	Apparent Cause	State	Failure Date	Date Posted
Buckeye Partners, LP	1845	HL	Incorrect Operation	PA	17-Jun-12	29-Mar-16
Buckeye Partners, LP	1845	HL	Corrosion Failure	PA	13-Jul-12	23-Nov-15
Magellan Pipeline Company, LP	26610	HL	Other Incident Cause	IA	22-Nov-12	9-Jun-16
Magellan Pipeline Company, LP	26610	HL	External Corrosion	KS	25-Nov-12	8-Jun-17
Buckeye Partners, LP	1845	HL	Outside Force Damage	NJ	10-Dec-12	28-Mar-16
CCPS Transportation, LLC	32080	HL	Equipment Failure	OK	14-Dec-12	21-Dec-15
Lion Oil Trading & Transportation, Inc.	11551	HL	Corrosion Internal	AR	9-Mar-13	4-Jan-17
Mobil Pipe Line Company	12628	HL	Material Failure - Pipe	AR	29-Mar-13	22-Jun-16
Enbridge Pipelines, LLC	31947	HL	Internal Corrosion	OK	17-May-13	24-Feb-14
Enterprise Products Operating, LLC	31618	HL	Material Failure Pipe	IL	12-Aug-13	12-Aug-13
Buckeye Partners, LP	1845	HL	Incorrect Operation	NJ	13-Aug-13	8-Oct-13
Buckeye Partners, LP	1845	HL	Other, Miscellaneous	NY	16-Oct-13	21-Sep-16
Columbia Gas Transmission Company	2620	NG	Natural Force Damage	KY	13-Feb-14	28-Mar-17
Williams Partners Operating LLC	39054	NG	Incorrect Operation	WA	31-Mar-14	29-Apr-16
Buckeye Partners, LP	1845	HL	Internal Corrosion	NJ	20-Aug-14	6-Jun-17
Enterprise Products Operating, LLC	31618	HL	Material Failure Pipe or Weld	WV	26-Jan-15	14-Apr-17
Plains Pipeline, LP	300	HL	Corrosion External	CA	19-May-15	1-Jun-16
Transcontinental Gas Pipeline Company	19570	GT	Material Failure - Cracking	PA	9-Jun-15	9-Jun-16
Centurion	31888	HL	Equipment Failure	TX	2-Aug-15	6-Jun-15
Tennessee Gas Pipeline Company	19160	GT	Corrosion External	TX	3-Aug-15	4-Jan-17
Columbia Gas Transmission, LLC	2616	GT	Equipment Failure	PA	9-Aug-15	22-Sep-16
Kiantone Pipeline Company	10250	HL	Natural Force Damage	PA	25-Aug-15	9-Jun-16

Operator Name	Operator Id	System Type	Apparent Cause	State	Failure Date	Date Posted
Gulf South Pipeline Company, LP	31728	GT	Other Incident Cause	LA	26-Aug-15	7-Jul-17
Gulf South Pipeline Company, LP	31728	GT	Other Incident Cause	LA	26-Aug-15	29-Jun-17
Enterprise Crude Pipeline	30829	HL Apparent	Tank Line Failure Due to Internal Corrosion	OK	1-Dec-15	19-Apr-17
TC Oil Pipeline Operations, Inc.	32334	HL	Material Failure	SD	16-Nov-17	20-Dec-18
Natural Gas Pipeline Co of America	13120	GT	Third-Party Damage	IL	5-Dec-17	9-Jul-18
Texas Gas Transmission LLC	19270	HL	Material Failure of Pipe or Weld	LA	9-Sept-15	22-Jun-17

Appendix A2 A summary of the incident investigation reports that was available for the study.

Reports #	Title	Report date	Incident date	Primary root cause	Commodity released	Useful pages	Used in NLP study
130804	Failure Investigation Report Enbridge Energy Limited Partnership Line 1 Leak, Equipment Failure	5/30/2015	7/28/2010	Leak	Crude oil	2	No
133500	Failure Investigation Report Sunoco Pipeline L.P. Darby Creek Tank Leak	7/12/2013	2/8/2011	Leak	Crude oil	3	Yes
144352	Failure Investigation Report Enterprise Products Operating, LLC Material Failure	5/26/2015	8/12/2013	Rupture, Material Failure	Ethane/Propane Mix	6	No
151195	Failure Investigation Report Kiantone Pipeline Company Cracked 2 inch NPS Drain - Crude Oil Leak, West Seneca Terminal, NY	3/28/2016	8/25/2015	Leak	Heavy Crude Oil	2	No
136756	Failure Investigation Report Belle Fourche, Sussex Diesel Line Release	10/8/2013	11/13/2011	Operator Error/Incorrect Operations	Diesel, Fuel Oil	6	No
129897	Failure Investigation Report Bridger Lake LLC Crude Oil Release	8/4/2011	4/2/2010	Rupture caused by Operator Error	Light Crude Oil	4	No

Reports #	Title	Report date	Incident date	Primary root cause	Commodity released	Useful pages	Used in NLP study
147585	Failure Investigation Report Buckeye Linden Station Internal Corrosion Leak, 8-inch Relief Line	2/11/2016	8/20/2014	Leak due to internal corrosion on dead leg station piping segment	#2 Diesel Fuel	3	Yes
139214	Failure Investigation Report Buckeye Partners LP, Turnpike Road NY Line 803 Excavation Damage	4/18/2012	9/20/2011	Leak caused by excavation damage (farmer using plow to install drain tile in field)	Gasoline	3	No
140298	Failure Investigation Report Buckeye Macungie Tank 230 bottom weld failure	7/17/2013	7/13/2012	Leak from crack in weld between tank bottom and wall	Hazardous Liquid (Gasoline)	2	No
141992	Failure Investigation Report Buckeye Partners Pipeline Gasoline Leak	7/25/2013	12/10/2012	Pinhole Leak / Other Outside Force Damage; Electrical Arcing from Other Equipment or Facility	Hazardous Liquid (Gasoline)	3	No

Reports #	Title	Report date	Incident date	Primary root cause	Commodity released	Useful pages	Used in NLP study
133920	Failure Investigation Report Buckeye External Corrosion Pit near Shippingport, PA	8/5/2011	3/20/2011	Pipeline Leak due to localized external corrosion pit	Diesel	2	No
142160	Failure Investigation Report Enbridge Pipeline Sump Pump Discharge Flex Hose Failure	7/16/2013	12/14/2012	Equipment Failure - Flex Hose Failure on Sump Pump Discharge Resulting in the Release of Crude Oil	Hazardous Liquid (Crude Oil)	5	No
151238	Failure Investigation Report Centurion Pipeline L.P.	5/13/2016	8/2/2015	Tank Mixer Failure	Crude Oil	7	No
130345	Failure Investigation Report Chevron Leak	4/14/2011	6/11/2010	Leak caused by Other Outside Force Damage – Electrical Arcing	Crude Oil	3	No
132398	Failure Investigation Report Chevron Pipeline Crude Oil Release	6/7/2011	12/1/2010	Leak caused by Inadequate Procedures for Draining Water	Crude Oil Condensate	12	No

Reports #	Title	Report date	Incident date	Primary root cause	Commodity released	Useful pages	Used in NLP study
133527	Chevron Pipeline Company, Grand Bay 10-inch Pipeline, Plaquemines Parish, Louisiana	5/30/2012	1/26/2011	Brittle, tensile fracture at pre-existing mechanical damage	Crude Oil	3	No
135347	Failure Investigation Report Central Florida Pipeline 10-inch Jet Fuel Pipeline Failure	10/12/2012	7/22/2011	Pipe leaked due to mechanical damage	Aviation jet fuel (Jet-A)	11	No
130575	Failure Investigation Report Dixie Pipeline Company 8-inch Propane Pipeline Release	8/31/2011	7/5/2010	Leak due to third party excavation damage	Propane (HVL)	6	No
143591	Failure Investigation Report Enbridge Pipelines, LLC, Tank 3013 24-inch Fill Line failure in Cushing, OK	2/24/2014	5/17/2013	Internal Corrosion, Microbiologically Influenced (MIC)	West Texas Intermediate Crude Oil	7	Yes
149469	Failure Investigation Report Enterprise Products Operating, LLC: ATEX Ethane Pipeline Failure, Follansbee, West Virginia	2/24/2016	1/26/2015	Girth Weld Failure Caused by Ductile Tensile Overload	Ethane	4	No
130938	Failure Investigation Report Enterprise Products	10/21/2011	8/27/2010	Circumferentially-oriented stress	Liquid Propane	3	Yes

Reports #	Title	Report date	Incident date	Primary root cause	Commodity released	Useful pages	Used in NLP study
	Propane Line Crack			corrosion cracking caused the pipe to separate			
137399	Failure Investigation Report Enterprise Products Pipeline Rio Grande PL Girth Weld Failure	9/12/2013	12/27/2011	Girth weld failure (complete separation of circumference of weld)	LPG Products (Propane/Butane)	5	No
151766	Failure Investigation Report Enterprise Crude Pipeline, LLC, Cushing West Tank Farm Release	12/27/2016	12/1/2015	Tank line failure due to internal corrosion	Crude Oil	5	Yes
133587	Failure Investigation Report Enterprise Cushing Terminal	5/30/2012	2/21/2011	Incorrect Operation	Crude Oil	5	No
139211	Failure Investigation Report Enterprise Crude Pipeline, LLC (Cushing West Tank Farm, Cushing, OK, Line C75)	2/3/2014	4/8/2012	Breakout tank line failure due to internal corrosion	Crude Oil	8	Yes
132719	Failure Investigation Report Denbury Green	10/1/2013	12/20/2010 and 02/14/2011	Small seam weld penetrators from manufacture of the pipe	CO2	7	No

Reports #	Title	Report date	Incident date	Primary root cause	Commodity released	Useful pages	Used in NLP study
135547	Failure Investigation Report Harbor Pipeline Fire Incident, Mansfield Township, NJ	7/5/2012	10/11/2010	Fire – Incorrect Operation	ULSD Diesel Fuel	2	No
142985	Failure Investigation Report Lion Oil Trading & Transportation, Inc. Suction Strainer Failure - Magnolia Tank Farm	9/12/2013	3/9/2013	Suction strainer failed resulting in the release of 5,600 bbl of crude oil	Crude Oil	5	No
147517	Failure Investigation Report Magellan Pipeline Company, LP External Corrosion, Crevice and Atmospheric	3/28/2015 (11/8/2016)	11/25/2012	Pinhole leak at bridge pipe support; crevice and atmospheric corrosion	Refined Product—Jet Fuel	6	Yes
130689	Failure Investigation Report Magellan Ammonia Line 501 Buckle	7/1/2011	7/23/2010	Leak, pipe buckle and crack resulting from compressive overload	Anhydrous Ammonia	5	No
136869	Failure Investigation Report Magellan Pipeline Company, Orion 20-inch Pipeline, 3012 Tank Line, East Houston Terminal	11/6/2012	12/1/2011	Operator Error/Incorrect Operation	Diesel	7	No

Reports #	Title	Report date	Incident date	Primary root cause	Commodity released	Useful pages	Used in NLP study
136157	Failure Investigation Report Magellan #6-10" Excavation Damage Lawrence, Kansas	11/29/2012	10/6/2011	Mechanical puncture of pipeline by third party excavator	Refined Product – Diesel Fuel	3	No
129379	Failure Investigation Report Mid-Valley Pipeline Internal Corrosion	7/11/2011	3/1/2010	Internal corrosion in manifold area of tank farm	Crude Oil	4	Yes
143154	Failure Investigation Report Mobil Pipe Line Company; Pegasus Pipeline, Mayflower, AR	10/23/2013	3/29/2013	ERW Seam Failure	Wabasca Heavy Crude Oil	12	No
150537	Failure Investigation Report Plains Pipeline, LP, Line 901 Crude Oil Release, May 19, 2015 Santa Barbara County, California	5/5/2016	5/19/2015	External Corrosion	Crude Oil	17	No
129735	Failure Investigation Report SFPP LP Bleed Fitting Corrosion	11/9/2010	3/16/2010	Leak from Bleed Fitting due to Internal Corrosion	Refined Products	2	Yes
135866	Failure Investigation Report Shell Houma to Houston (Ho-Ho) Pipeline	6/29/2012	11/16/2010	Corrosion Fatigue Cracking	Crude Oil	5	Yes

Reports #	Title	Report date	Incident date	Primary root cause	Commodity released	Useful pages	Used in NLP study
130287	Failure Investigation Report Suncor Energy Pipeline Company (Suncor) Tank # 1168 Overfill	4/13/2012	6/14/2010	Break Out Tank Overflow	Crude Oil	3	No
129572	Failure Investigation Report Sunoco R&M Flange Gasket	4/28/2011	3/25/2010	Flange Leak caused by deteriorated gasket. The loss of pipe support and leakage through a closed valve contributed to the failure	Vacuum Gas Oil and Light Cycle Oil	2	No
158348	Failure Investigation Report Material Failure – Mechanical Damage from Original Construction – TC Oil Pipeline Operations, Inc	11/28/2018	11/16/2017	Rupture – Material Failure – Damage from Original Construction	Crude Oil	13	No
130425	Failure Investigation Report Whitecap (Chevron), 18” Offshore Failure	6/16/2011	3/25/2010	Leak/Outside force damage from contact with other pipeline	Crude Oil	4	No

REFERENCE

Abraham, W., Nichols, Sorrels, Agosto & Aziz. (2019). "Large Pipeline Explosion and Oil Spill Accidents Around the World." Retrieved 12/3/2019, 2019, from https://www.abrahamwatkins.com/Petrochemical-Accidents/Major-Pipeline-Explosions-Oil-Spills/.

Adedigba, S. A., F. Khan and M. Yang (2016). "Process accident model considering dependency among contributory factors." Process Safety Environmental Protection **102**: 633-647.

Allahyari, M., S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. Kochut (2017). "A brief survey of text mining: Classification, clustering and extraction techniques." arXiv preprint arXiv:.02919.

Allahyari, M., S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. J. a. p. a. Kochut (2017). "A brief survey of text mining: Classification, clustering and extraction techniques."

Allison, E. and B. Mandler (2018). Transportation of Oil, Gas, and Refined Products. Petroleum and the Environment, American Geosciences Institute.

American Petroleum Institute and Associattion of Oil Pipelines (2019). Pipeline Safety Excellence Performance 2019 Annual Liquids Report, American Petroleum Institute (API) & Associattion of Oil Pipelines (AOPL)

Amigó, E., J. Gonzalo, J. Artiles and F. Verdejo (2009). "A comparison of extrinsic clustering evaluation metrics based on formal constraints." Information retrieval **12**(4): 461-486.

Andersen, T. and A. Misund (1983). "Pipeline reliability: an investigation of pipeline failure characteristics and analysis of pipeline failure rates for submarine and cross-country pipelines." Journal of Petroleum Technology **35**(04): 709-717.

Awalt, J. (2019). P&GJ's Global Pipeline Construction Outlook. Pipeline and Gas Journal, Gulf Energy Information. **January**.

Banimostafa, A., S. Papadokonstantakis and K. Hungerbühler (2012). "Evaluation of EHS hazard and sustainability metrics during early process design stages using principal component analysis." Process safety environmental protection **90**(1): 8-26.

Baybutt, P. (2016). "Insights into process safety incidents from an analysis of CSB investigations." Journal of loss prevention in the process industries **43**: 537-548.

Bersani, C., L. Citro, R. V. Gagliardi, R. Sacile and A. M. Tomasoni (2010). "Accident occurrance evaluation in the pipeline transport of dangerous goods." Chemical Engineering Transactions: 249-254.

Bersani, C., L. Citro, R. V. Gagliardi, R. Sacile and A. M. J. C. E. T. Tomasoni (2010). "Accident occurrance evaluation in the pipeline transport of dangerous goods." 249-254.

Bholowalia, P. and A. Kumar (2014). "EBK-means: A clustering technique based on elbow method and k-means in WSN." International Journal of Computer Applications **105**(9).

Bolt, R., A. Hilgenstock, C. Kolovich, D. Velez Vega, A. Cappanera and O. Rasmussen (2006). A guideline: using or creating incident databases for natural gas transmission pipelines. International Pipeline Conference.

Breton, T., J. Sanchez-Gheno, J. Alamilla and J. Alvarez-Ramirez (2010). "Identification of failure type in corroded pipelines: A Bayesian probabilistic approach." Journal of hazardous materials **179**(1-3): 628-634.

Bubbico, R. (2018). "A statistical analysis of causes and consequences of the release of hazardous materials from pipelines." Journal of Loss Prevention in the Process Industries **56**: 458-466.

Bubbico, R. (2018). "A statistical analysis of causes and consequences of the release of hazardous materials from pipelines. The influence of layout." Journal of Loss Prevention in the Process Industries **56**: 458-466.

Canada Energy Regulator. (2019). "Incident Data." Retrieved 3.31.2019, 2019, from https://www.cer-rec.gc.ca/sftnvrnmnt/sft/dshbrd/mp/dt-eng.html.

Carpenter, P., M. Henrie, Y. Okamoto and P. Liddell (2019). Analysis of PHMSA Spill Data for Pipeline Spill Risk Analysis. PSIG Annual Meeting, OnePetro.
Carvalho, A., J. Rebello, M. Souza, L. Sagrilo and S. Soares (2008). "Reliability of non-destructive test techniques in the inspection of pipelines used in the oil industry." International journal of pressure vessels piping **85**(11): 745-751.
Center for Chemical Process Safety (2011). Process safety leading and lagging metrics. New York, AIChE.
Center for Chemical Process Safety (2019). Guidelines for Investigating Process safety Incidents. New York, AIChE.
Central Intelligence Agency. (2019). "The world factbook: pipelines." Retrieved 6/3/2019, 2019, from https://www.cia.gov/library/publications/the-world-factbook/fields/383.html.
Chokor, A., H. Naganathan, W. K. Chong and M. El Asmar (2016). "Analyzing Arizona OSHA injury reports using unsupervised machine learning." Procedia Engineering **145**: 1588-1593.
Cortes, C., X. Gonzalvo, V. Kuznetsov, M. Mohri and S. Yang (2016). "Adanet: Adaptive structural learning of artificial neural networks." arXiv preprint arXiv:1607.01097.
Cunha, S. B. (2012). Comparison and analysis of pipeline failure statistics. 2012 9th International Pipeline Conference, American Society of Mechanical Engineers Digital Collection.
Davis, P., J. Dubois, A. Olcese, F. Uhlig, J. Larivé and D. Martin (2006). "Performance of European cross-country oil pipelines." Statistical summary of reported spillages **54**.
Dey, P. K., S. O. Ogunlana and S. Naksuksakul (2004). "Risk-based maintenance model for offshore oil and gas pipelines: a case study." Journal of Quality in Maintenance Engineering.
El-Abbasy, M. S., A. Senouci, T. Zayed, F. Mirahadi and L. Parvizsedghy (2014). "Artificial neural network models for predicting condition of offshore oil and gas pipelines." Automation in Construction **45**: 50-65.
Esmaeili, B. and M. Hallowell (2012). Attribute-based risk model for measuring safety risk of struck-by accidents. Construction Research Congress 2012: construction challenges in a flat world.
European Gas Pipeline Incident Data Group (2018). Gas Pipeline Incidents 10th Report of the European Gas Pipeline Incident Data Group (period 1970 – 2016), European Gas Pipeline Incident Data Group,.
Ferjencik, M. (2011). "An integrated approach to the analysis of incident causes." Safety Science **49**(6): 886-905.
Ferjencik, M. (2014). "IPICA_Lite—Improvements to root cause analysis." Reliability Engineering System Safety **131**: 1-13.
Ferret, O. (2004). Discovering word senses from a network of lexical cooccurrences. COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics.
Fruchterman, T. M. and E. M. Reingold (1991). "Graph drawing by force-directed placement." Software: Practice experience **21**(11): 1129-1164.
Goh, Y. M. and C. Ubeynarayana (2017). "Construction accident narrative classification: An evaluation of text mining techniques." Accident Analysis Prevention **108**: 122-130.
Guo, Y., X. Meng, D. Wang, T. Meng, S. Liu and R. He (2016). "Comprehensive risk evaluation of long-distance oil and gas transportation pipelines using a fuzzy Petri net model." Journal of Natural Gas Science Engineering **33**: 18-29.
Halim, S. Z., S. Janardanan, T. Flechas and M. S. Mannan (2018). "In search of causes behind offshore incidents: Fire in offshore oil and gas facilities." Journal of Loss Prevention in the Process Industries **54**: 254-265.
Halim, S. Z., N. Quddus and H. Pasman (2021). "Time-trend analysis of offshore fire incidents using nonhomogeneous Poisson process through Bayesian inference." Process Safety Environmental Protection **147**: 421-429.
Halim, S. Z., M. Yu, H. Escobar and N. Quddus (2020). "Towards a causal model from pipeline incident data analysis." Process Safety Environmental Protection **143**: 348-360.

Halkidi, M., Y. Batistakis and M. Vazirgiannis (2001). "On clustering validation techniques." Journal of intelligent information systems **17**(2): 107-145.
Han, Z. and W. Weng (2011). "Comparison study on qualitative and quantitative risk assessment methods for urban natural gas pipeline network." Journal of hazardous materials **189**(1-2): 509-518.
He, X., X. Du, X. Wang, F. Tian, J. Tang and T.-S. Chua (2018). "Outer product-based neural collaborative filtering." arXiv preprint arXiv:.03912.
Higuchi, K. (2016). KH Coder 3 reference manual, Kioto: Ritsumeikan University.
Hollnagel, E. (2017). FRAM: the functional resonance analysis method: modelling complex socio-technical systems, CRC Press.
International Energy Agency. (2019). "Global energy demand rose by 2.3% in 2018, its fastest pace in the last decade." Retrieved 6/3/2019, 2019, from https://www.iea.org/newsroom/news/2019/march/global-energy-demand-rose-by-23-in-2018-its-fastest-pace-in-the-last-decade.html.
Jones, K. S. (1972). "A statistical interpretation of term specificity and its application in retrieval." Journal of documentation.
Kalantarnia, M., F. Khan and K. Hawboldt (2009). "Dynamic risk assessment using failure assessment and Bayesian theory." Journal of Loss Prevention in the Process Industries **22**(5): 600-606.
Kaszniak, M. (2010). "Oversights and omissions in process hazard analyses: Lessons learned from CSB investigations." Process Safety Progress **29**(3): 264-269.
Kelly, D. (2007). Bayesian Modeling of Time Trends in Component Reliability Data Via Markov Chain Monte Carlo Simulation, Idaho National Laboratory (INL).
Lam, C. and W. Zhou (2016). "Statistical analyses of incidents on onshore gas transmission pipelines based on PHMSA database." International Journal of Pressure Vessels Piping **145**: 29-40.
Lam, C., W. Zhou and Piping (2016). "Statistical analyses of incidents on onshore gas transmission pipelines based on PHMSA database." International Journal of Pressure Vessels **145**: 29-40.
Lappas, G. (2007). Estimating the size of neural networks from the number of available training data. International Conference on Artificial Neural Networks, Springer.
Leveson, N. G. (2016). Engineering a safer world: Systems thinking applied to safety, The MIT Press.
Li, X., G. Chen and H. Zhu (2016). "Quantitative risk analysis on leakage failure of submarine oil and gas pipelines using Bayesian network." Process Safety Environmental Protection **103**: 163-173.
Liu, S. J., S. L. Li, M. Jiang and D. He (2017). Quantitative identification of pipeline crack based on BP neural network. Key Engineering Materials, Trans Tech Publ.
Loper, E. and S. Bird (2002). "Nltk: The natural language toolkit." arXiv preprint cs/0205028.
Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard and D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations.
Manning, C. D. and H. Schütze (1999). Foundations of statistical natural language processing, MIT press.
Mazumder, R. K., A. M. Salman and Y. Li (2021). "Failure risk analysis of pipelines using data-driven machine learning algorithms." Structural Safety **89**: 102047.
Meng, X., G. Chen, G. Zhu and Y. Zhu (2019). "Dynamic quantitative risk assessment of accidents induced by leakage on offshore platforms using DEMATEL-BN." International Journal of Naval Architecture Ocean Engineering **11**(1): 22-32.
Mihalcea, R. and P. Tarau (2004). Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing.
Miner, G., J. Elder IV, A. Fast, T. Hill, R. Nisbet and D. Delen (2012). Practical text mining and statistical analysis for non-structured text data applications, Academic Press.
Muhlbauer, W. K. (2004). Pipeline risk management manual: ideas, techniques, and resources, Elsevier.
Naghavi-Konjin, Z., S.-B. Mortazavi, H. Asilian-Mahabadi and E. Hajizadeh (2020). "Ranking the occupational incident contributory factors: a Bayesian network model for the petroleum industry." Process Safety Environmental Protection **137**: 352-357.

Najafi, M. and G. Kulandaivel (2005). Pipeline condition prediction using neural network models. Pipelines 2005: Optimizing Pipeline Design, Operations, and Maintenance in Today's Economy**:** 767-781.
Nakata, T. (2017). Text-mining on incident reports to find knowledge on industrial safety. 2017 Annual Reliability and Maintainability Symposium (RAMS), IEEE.
National Energy Board (2019). Incident Data: Methodology, National Energy Board, Canada.
National Transportation Safety Board (2011). Pacific Gas and Electric Company Natural Gas Transmission Pipeline Rupture and Fire San Bruno, California September 9, 2010 National Transportation Safety Board
National Transportation Safety Board (2012). Enbridge Incorporated Hazardous Liquid Pipeline Rupture and Release Marshall, Michigan July 25, 2010, National Transportation Safety Board.
National Transportation Safety Board (2018). Pipeline Accident Brief TransCanada Corporation Pipeline (Keystone Pipeline) Rupture, Amherst, South Dakota, National Transportation Safety Board.
Nuchitprasittichai, A. and S. Cremaschi (2013). "An algorithm to determine sample sizes for optimization with artificial neural networks." AIChE Journal **59**(3): 805-812.
Occupational Safety and Health Administration (2015). Incident (accident) investigations: A guide for employers, Occupational Safety and Health Administration.
Ochiai, K. and S. Usui (1993). Improved kick out learning algorithm with delta-bar-delta-bar rule. IEEE International Conference on Neural Networks, IEEE.
Omidi, L., S. A. Zakerian, J. N. Saraji, E. Hadavandi and M. S. Yekaninejad (2018). "Safety performance assessment among control room operators based on feature extraction and genetic fuzzy system in the process industry." Process Safety Environmental Protection **116**: 590-602.
Oyedele, A., A. Ajayi, L. Oyedele, J. M. D. Delgado, L. Akanbi, O. Akinade, H. Owolabi and M. Bilal (2021). "Deep learning and Boosted trees for injuries prediction in power infrastructure projects." Applied Soft Computing: 107587.
Paltrinieri, N., G. Scarponi, F. Khan and S. Hauge (2014). "Addressing dynamic risk in the petroleum industry by means of innovative analysis solutions." Chemical Engineering Transactions **36**: 451-456.
Paltrinieri, N., A. Tugnoli, J. Buston, M. Wardman and V. Cozzani (2013). "Dynamic procedure for atypical scenarios identification (DyPASI): a new systematic HAZID tool." Journal of Loss Prevention in the Process Industries **26**(4): 683-695.
Papadakis, G. A. (1999). "Major hazard pipelines: a comparative study of onshore transmission accidents." Journal of Loss Prevention in the Process Industries **12**(1): 91-107.
Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg (2011). "Scikit-learn: Machine learning in Python." the Journal of machine Learning research **12**: 2825-2830.
Perrow, C. (2011). Normal accidents, Princeton university press.
Pipeline and Hazardous Materials Safety Administration. (2019). "Pipeline Incident 20 Year Trends." Retrieved 6/3/2019, 2019, from https://www.phmsa.dot.gov/data-and-statistics/pipeline/pipeline-incident-20-year-trends.
Pipeline and Hazardous Materials Safety Administration. (2019). "Pipeline Incident Flagged Files." Retrieved 5.31.2019, 2019, from https://www.phmsa.dot.gov/data-and-statistics/pipeline/pipeline-incident-flagged-files.
Pyun, H., K. Kim, D. Ha, C.-J. Lee and W. B. Lee (2020). "Root causality analysis at early abnormal stage using principal component analysis and multivariate Granger causality." Process Safety Environmental Protection **135**: 113-125.
Quddus, N., M. Yu, N. Tamim, S. Rahmani and M. S. Mannan (2018). "Risk assessment of class 3 (PG II & III) hazardous materials in transportation." Process Safety Progress **37**(3): 376-381.
Ramírez-Camacho, J. G., F. Carbone, E. Pastor, R. Bubbico and J. Casal (2017). "Assessing the consequences of pipeline accidents to support land-use planning." Safety science **97**: 34-42.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning, Piscataway, NJ.
Rathnayaka, S., F. Khan and P. Amyotte (2011). "SHIPP methodology: Predictive accident modeling approach. Part I: Methodology and model description." Process safety environmental protection **89**(3): 151-164.
Rausand, M. and A. Hoyland (2003). System reliability theory: models, statistical methods, and applications, John Wiley & Sons.
Restrepo, C. E., J. S. Simonoff and R. Zimmerman (2009). "Causes, cost consequences, and risk implications of accidents in US hazardous liquid pipeline infrastructure." International Journal of Critical Infrastructure Protection **2**(1-2): 38-50.
Robinson, S. (2019). "Temporal topic modeling applied to aviation safety reports: A subject matter expert review." Safety science **116**: 275-286.
Robinson, S., W. Irwin, T. Kelly and X. Wu (2015). "Application of machine learning to mapping primary causal factors in self reported safety narratives." Safety science **75**: 118-129.
Rodionov, A., D. Kelly and J. Uwe-Klügel (2009). "Guidelines for analysis of data related to ageing of nuclear power plant components and systems." JRC Scientific Technical Reports, EUR **23954**: 88-90.
Romesburg, C. (2004). Cluster analysis for researchers, Lulu. com.
Salton, G. and C. Buckley (1988). "Term-weighting approaches in automatic text retrieval." Information processing management **24**(5): 513-523.
Senouci, A., M. S. El-Abbasy and T. Zayed (2014). "Fuzzy-based model for predicting failure of oil pipelines." Journal of Infrastructure Systems **20**(4): 04014018.
Senouci, A., M. Elabbasy, E. Elwakil, B. Abdrabou and T. Zayed (2014). "A model for predicting failure of oil pipelines." Structure Infrastructure Engineering **10**(3): 375-387.
Shan, K., J. Shuai, K. Xu and W. Zheng (2018). "Failure probability assessment of gas transmission pipelines based on historical failure-related data and modification factors." Journal of Natural Gas Science Engineering **52**: 356-366.
Sidarta, D. E., J. Kyoung, J. O'Sullivan and K. F. Lambrakos (2017). Prediction of offshore platform mooring line tensions using artificial neural network. International Conference on Offshore Mechanics and Arctic Engineering, American Society of Mechanical Engineers.
Siler-Evans, K., A. Hanson, C. Sunday, N. Leonard and M. Tumminello (2014). "Analysis of pipeline accidents in the United States from 1968 to 2009." International journal of critical infrastructure protection **7**(4): 257-269.
Single, J. I., J. Schmidt and J. Denecke (2020). "Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing." Safety Science **129**: 104747.
Srinivasan, R. and N. T. Nhan (2008). "A statistical approach for evaluating inherent benign-ness of chemical process routes in early design stages." Process Safety Environmental Protection **86**(3): 163-174.
Syeda, K. N., S. N. Shirazi, S. A. A. Naqvi and H. J. Parkinson (2017). "Exploiting Natural Language Processing for Analysing Railway Incident Reports." IGI Glob. Publ.: 1-18.
Tanguy, L., N. Tulechki, A. Urieli, E. Hermann and C. Raynal (2016). "Natural language processing for aviation safety reports: from classification to interactive analysis." Computers in Industry **78**: 80-95.
Taylor, J. (2017). "Automated HAZOP revisited." Process Safety Environmental Protection **111**: 635-651.
Taylor, R. (2016). "Can process plant QRA reduce risk?–experience of ALARP from 92 QRA studies over 36 years." Chemical Engineering Transactions **48**: 811-816.
Tixier, A. J.-P., M. R. Hallowell, B. Rajagopalan and D. Bowman (2016). "Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports." Automation in Construction **62**: 45-56.
Toman, M., R. Tesar and K. Jezek (2006). "Influence of word normalization on text classification." Proceedings of InSciT **4**: 354-358.

Tulechki, N. (2015). Natural language processing of incident and accident reports: application to risk management in civil aviation.
Van der Maaten, L. and G. Hinton (2008). "Visualizing data using t-SNE." Journal of machine learning research **9**(11).
Veling, A. and P. Van Der Weerd (1999). Conceptual grouping in word co-occurrence networks. IJCAI.
Verma, A. and J. Maiti (2018). "Text-document clustering-based cause and effect analysis methodology for steel plant incident data." International journal of injury control safety promotion **25**(4): 416-426.
Wagstaff, K., C. Cardie, S. Rogers and S. Schrödl (2001). Constrained k-means clustering with background knowledge. Icml.
Wang, H. and I. J. Duncan (2014). "Likelihood, causes, and consequences of focused leakage and rupture of US natural gas transmission pipelines." Journal of loss prevention in the process industries **30**: 177-187.
Wikipedia. (2019). "List of countries by total length of pipelines." Retrieved 6/3/2019, 2019, from https://en.wikipedia.org/wiki/List_of_countries_by_total_length_of_pipelines.
Wikipedia. (2019). "List of pipeline accidents." Retrieved 6.3.2019, 2019, from https://en.wikipedia.org/wiki/List_of_pipeline_accidents.
Wu, J., R. Zhou, S. Xu and Z. J. J. o. L. P. i. t. P. I. Wu (2017). "Probabilistic analysis of natural gas pipeline network accident based on Bayesian network." **46**: 126-136.
Xin, P., F. Khan and S. Ahmed (2017). "Dynamic hazard identification and scenario mapping using Bayesian network." Process Safety Environmental Protection **105**: 143-155.
Xu, W.-Z., C. B. Li, J. Choung and J.-M. Lee (2017). "Corroded pipeline failure analysis using artificial neural network scheme." Advances in engineering software **112**: 255-266.
Yegnanarayana, B. (2009). Artificial neural networks, PHI Learning Pvt. Ltd.
Yu, M., N. Quddus, S. C. Peres, S. Sachdeva and M. S. Mannan (2017). "Development of a safety management system (SMS) for drilling and servicing operations within OSHA jurisdiction area of Texas." Journal of Loss Prevention in the Process Industries **50**: 266-274.
Zangenehmadar, Z. and O. Moselhi (2016). "Assessment of remaining useful life of pipelines using different artificial neural networks models." Journal of performance of constructed facilities **30**(5): 04016032.
Zhang, F., H. Fleyeh, X. Wang and M. Lu (2019). "Construction site accident analysis using text mining and natural language processing techniques." Automation in Construction **99**: 238-248.
Zhang, J., J. Fu, H. Hao, G. Fu, F. Nie and W. Zhang (2020). "Root causes of coal mine accidents: Characteristics of safety culture deficiencies based on accident statistics." Process Safety Environmental Protection **136**: 78-91.
Zhang, Z., P. Zweigenbaum and R. Yin (2018). Efficient generation and processing of word co-occurrence networks using corpus2graph. Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12).